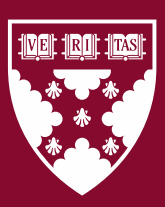
Working Paper 22-001

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

Annamaria Conti
Christian Peukert
Maria Roche



Harvard Business School

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

Annamaria Conti
IE Business School

Christian Peukert
HEC Lausanne

Maria Roche
Harvard Business School

Working Paper 22-001

Copyright © 2021, 2022, and 2023 by Annamaria Conti, Christian Peukert, and Maria Roche.

Working papers are in draft form. This working paper is distributed for purposes of comment and discussion only. It may not be reproduced without permission of the copyright holder. Copies of working papers are available from the author.

We thank Kimon Protopapas, Ilia Azizi, and Umar Ali for excellent research assistance and Jorge Guzman for sharing data from Guzman and Li (2021). 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.

Funding for this research was provided in part by Harvard Business School.

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

August 9, 2023

Abstract

We study the participation of nascent firms in open source communities and its implications for attracting funding. To do so, we link data on 160,065 U.S. startups from Crunchbase to their activities on the open source development platform GitHub. In a matched sample of firms with and without GitHub activities, difference-in-differences models reveal a substantial increase in the likelihood of being funded after firms engage with open source communities on GitHub. This relationship appears especially strong for firms working on novel technologies and negative for those facing high levels of competition, highlighting potential differences in opportunity costs of engagement. To provide insight regarding mechanisms, we classify startups' technology use-cases on GitHub using machine learning and exploit data on product launches. Our results from these additional analyses support the notion that one important channel driving our results is the provision of access to external knowledge for technology improvement. Open source communities may thereby help startups improve their technologies to create a, at least, minimum viable product.

Keywords: Startups, Knowledge, Open Source, GitHub, Machine Learning, Venture Capital

*Conti: annamaria.conti@ie.edu, Peukert: christian.peukert@unil.ch, Roche: mroche@hbs.edu. We thank Kimon Protopapas, Ilia Azizi, and Umar Ali for excellent research assistance and Jorge Guzman for sharing data from Guzman and Li (2021). Many thanks to Bruno Cassiman, Matt Higgins, David Hsu, Celine Fei, Harsh Ketkar, Tobias Kretschmer, Rem Koning, Frank Nagle, Melissa Perri, and Toby Stuart for advice on drafts. This manuscript benefited from many helpful comments provided at CEAR, 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

> *"There is a huge open source community of code and developers to work with and draw inspiration from on GitHub. To build better developer tools, we need a direct line to it. GitHub accelerates production, progress, and connections, bringing us closer to our users." (Developer Advocate, Stripe)*[1]

Open source communities (OSCs) have been increasing in importance as suppliers of knowledge (Dahlander, 2005). In fact, recent survey results suggest that firms rely twice as much on open source for external supply of knowledge than patents.[2] Provided how much firms and society depend on open source (Greenstein and Nagle, 2014), understanding the dynamics of open source engagement 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) or how its use relates to business models (Dahlander and Magnusson, 2008) rather than how its use may impact achieving important performance milestones.

In this paper, we examine if interacting with OSCs can help nascent firms 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). 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, where the knowledge pieces used to build technologies come from still largely remains an open question.

We contribute new empirical findings on the relationship between nascent firms' engagement with OSCs and raising funds. To do so, we collect a novel dataset that combines firm, technology, and performance information. First, we exploit data encompassing 160,065 U.S. startups listed on Crunchbase that were founded between 2005 and 2020 as our initial sample. We also access information on startup activities on the GitHub open source platform, which describes itself as the

[1] `https://github.com/enterprise/startups`, accessed March 1, 2023.

[2] For example, data from a representative survey of about 5000 German firms (Mannheim Innovation Panel), shows that close to 20% of firms use open source to access external knowledge, while only 10% use patents of others to do so. See Figure A1 for details.

place "where the world builds software". We then link firms to organization accounts on GitHub and access the public record of their activities on GitHub from GHArchive, a community-led project that logs GitHub since 2011. We additionally combine these data with information on startup product launches available from Product Hunt, a repository of technology product launches that started in 2014 and was recently acquired by AngelList. Product Hunt has become the website where startups find information about existing products and venture capitalists (VCs) scout investment opportunities. The combination of these three 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 – if a firm engages on the platform –, and the technologies and products it develops.

The first question we address is whether interacting with OSCs makes startups more attractive to potential investors. For this scope, we estimate a difference-in-differences model, assessing how the likelihood of being funded changes after startups become engaged with external repositories[3] on GitHub, relative to a matched sample of startups founded in the same year and location as the treated startups, and developing comparable technologies, with similar human capital. We saturate this model with a host of fixed effects to hold constant fixed differences among startups and to account for technology shocks and life cycle differences. Our results suggest that engaging with OSCs on GitHub is associated with an increase in the likelihood of receiving funding by at least 36%. These findings are corroborated by the exploitation of a quasi-experimental change in the relative cost of engaging with external repositories. Our results further reveal that the relationship is strongest when startups are developing novel technologies, for which the benefits of relying on input from the open source may offset appropriability concerns. Conversely, the relationship is weakest when the level of competition is high, and, thus, appropriability concerns are potentially more relevant. We consistently observe that although startups operating in the Platform domain have the highest propensity to be active on GitHub, Artificial Intelligence (AI) & Blockchain startups benefit the most in attracting funds. As the latter startups develop more novel and complex technologies (Jordan and Mitchell, 2015), open source might help them bring these technologies to the market faster.

[3]In this paper, we make the distinction between activities related to internal – a focal company owns – and external – a focal company does not own – repositories that can be publicly observed on *github.com*.

We further provide suggestive evidence that our results are unlikely only driven by technology development life cycles. Specifically, we compare the effect of startups engaging with external repositories on attracting funds with the impact of the startups' engagement with their internal GitHub repositories. We show very pronounced differences between these relationships, which we should not have observed had life cycle effects been prominent.

Overall, we are able to empirically establish a robust relationship between startups' engagement with OSCs on GitHub and achieving funding milestones. However, it is possible that startups use GitHub only as an endorsing entity to increase visibility (Rysman and Simcoe, 2008), as has been shown in other "closed" contexts (Conti et al., 2013a,b; Hsu and Ziedonis, 2013), rather than to also improve their technology. To understand the potential contribution of engagement with OSCs to the technology development of a startup, we distinguish between different technology use cases that startups interact with on GitHub. Using natural language processing and machine learning methods (Miric et al., 2022), we classify the external repositories that startups interact with according to the following use cases: Software Development/Backend (SD/BE), Machine Learning (ML), Application Programming Interface (API), and User Interface (UI). We find that startups derive positive gains by interacting with all types of external repositories though there are important differences across domains. The large spectrum of technology use cases which startups interact with on GitHub provides a first indication that engaging with OSCs may act as a critical source of relevant knowledge.

To further probe into the mechanisms through which engagement with OSCs on GitHub may be helpful for startups to attract funds, we assess investor response to the type of output startups generate on GitHub. We show that investors do not react to "cosmetic" changes startups add to the documentation of their repositories ("readme files"). As such, this finding provides further support for a knowledge production explanation by which open source does not merely act as an amplifier of visibility.

To bring our exploration full circle, we further examine whether engaging with OSCs is correlated with other palpable measures of startup technology production, such as launching new products. Our results using product launches on Product Hunt suggest that engaging with OSCs on GitHub can contribute to completing, at least, a minimum viable product. On the whole, our results support the notion that startups are "beefing up" their technologies before they receive early-stage financing – by engaging with external knowledge from OSCs rather than merely gaining visibility, though the

latter may play a role. In closing our investigation, we show that the relationship we have uncovered is especially strong when startups attract funds from VCs and successful investors. Such investors have been shown by the literature to be particularly inclined to value a startup's technology over other aspects (Conti et al., 2013b).

2 Guiding framework

Innovation is widely understood as a process of recombinant search whereby innovators experiment with new components or new combinations of existing components (Fleming, 2001). Research suggests that innovating in isolation is extremely difficult (Laursen and Salter, 2006) and that embracing an "open innovation paradigm" using internal and external knowledge components is crucial for firms to advance their technology pipeline (Chesbrough, 2006; Grimpe and Kaiser, 2010). This may be particularly the case for novel technologies, whose development may require solving relatively complex problems. While engagement with OSCs can increase the number of knowledge components available and the variability of these components – both of which are fundamental for producing breakthrough innovations – this approach is not exempt from drawbacks (Almirall and Casadesus-Masanell, 2010; Dahlander et al., 2016). In fact, relative to a hierarchical organization with direct internal chains of command, engaging with external knowledge can heighten coordination costs (Greenstein, 1996) and give rise to appropriability concerns (Buss and Peukert, 2015; Laursen and Salter, 2014).

The tradeoff between adoption and appropriability is particularly relevant for technology startups (Gans and Stern, 2003). Because of their youth and small size, startups are more likely to both gain by drawing from external sources of ideas and be exposed to appropriability problems. This tradeoff may further differ depending on the type of technology a startup is developing and under what type of market conditions it operates. For example, especially novel technology may depend on co-development efforts with other organizations to create a functioning product, whereas startups entering a fairly competitive market may be more concerned about protecting their intellectual property to achieve competitive advantage. Thus, it remains an open question whether and to what extent engagement with external knowledge through engagement with OSCs can improve the innovation pipeline of nascent firms.

Such improvement in the technology of a startup is, however, critical for these firms to be able to

attract funds. As the existing literature has highlighted, in addition to the team, the underlying technology of a venture represents an important predictor of obtaining financing (Bernstein et al., 2017; Gompers et al., 2020; Kaplan et al., 2009; Roche et al., 2020b). Raising funds is widely considered to be a fundamental milestone for venture growth.

Beyond enabling a startup to improve its technology, OSCs may serve the role of an endorsing entity (Rysman and Simcoe, 2008). Startups may thereby rely on engagement with OSCs to gain visibility in the market for funding, by signaling their technology to potential investors (Conti et al., 2013a,b; Hsu and Ziedonis, 2013). Improving a startup's innovation stack and gaining visibility need not be mutually exclusive goals. As we know from Spence (2002), signals, such as education, may both improve productivity and allow highly productive individuals to separate themselves from less productive ones.

From this, the goals of this paper are four-fold. First, we aim to provide empirical evidence on the nature of the relationship between startup engagement with OSCs and achieving funding milestones. Second, we set out to unveil heterogeneity among different types of technologies that may benefit or even be harmed by such engagement. Third, our goal is to further our understanding of the underlying mechanism, namely, whether startups participate in OSCs to upgrade their technology or just to increase their visibility. Finally, we provide some indication of which investors are most sensitive to these activities.

3 Data

To build our dataset, we combine data on U.S. startups and their investors from Crunchbase with information on startups' GitHub activities available from GHArchive and the GitHub API and with information on technology products listed on Product Hunt.

3.1 Crunchbase

Crunchbase is an online directory that records fine-grained information on a large sample of startups, their founders, and their investors. As described in 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 into 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 extract information pertaining to all the recorded U.S. startups that were founded between 2005 and 2020. This amounts to 160,065 startups, for which we have data encompassing their founding dates, industry group 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.

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 develop a measure of how much a startup's technology is related to software. This measure 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 ⟩

3.2 GitHub Activities

GitHub is our next source of information. This is a hosting service for software development and collaborative version control. With 40 million public repositories in April 2021, GitHub was the largest host of source code[5] at the time of this study and has come to be known as the place "where the world builds software" (*GitHub.com*). Individuals and organizations use GitHub to host their own projects and upgrade them by using code and information from other existing projects as well as to contribute to the projects of others. 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.[6]

[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 August 4, 2023.

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

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

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 containing projects. The underlying version control system, Git, effectively stores historical versions 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 most frequently used way to interact with the repositories of other users is called *forking*. By 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. Forking plays a significant role in how knowledge is shared, created, and how collaboration occurs in OSCs.

We use the GitHub API to collect all organization accounts on GitHub. From these accounts, we extract the websites of the account owners. We use this information to link GitHub organization accounts to 14,881 Crunchbase company profiles. Over 60% of the startups with a GitHub organization account are described by Crunchbase's industry group keyword "software", 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 community-led project 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.

As reported in Table 1a, 9% of the startups in our sample have an organization account on GitHub, while 8% have and organization account and have engaged in an external activity on GitHub, namely interacted with an external repository the startups do not directly control. The latter serves as our primary measure of *engagement with OSCs*. Among the events through which startups engage with external repositories, forks are the largest portion (96%), followed by watch events (2%) and push events (1%).[7] This distribution suggests that startups may engage with external repositories mostly to develop their own technologies and less so to provide comments or contribute to other users' projects. As reported in Figure 1, there is a stark increase in the number of startups' engagements with external repositories over time, which tracks the increase in startups' overall activities on GitHub.

[7] A watch event is the bookmarking ("starring") of another user's repository. A push event is the "pushing" of commits to a given repository.

⟨ Insert Figure 1 about here ⟩

As suggested by recent work, machine learning 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 external repositories with which the organizations interacted through commits, pull requests or forks according to their type. We identify repositories that pertain to SD/BE, ML, API, and UI. By doing so, we consider a comprehensive set of use cases that are relevant for the development of digital technologies (Yoo et al., 2012).[8]

We report a descriptive representation of the output obtained from the algorithm in Figure A2, displaying the most common words for each of the categories we consider. Additionally, we report in Table 1b descriptive statistics summarizing startups' external activities on GitHub. Here, we show that startups prevalently engage with external repositories related to SD/BE and API.

3.3 Product Hunt

We use data from Product Hunt to assess how a startup's activities on GitHub relate to the development of products. Product Hunt is a website where individuals share new products. The venture was founded in 2013 by U.S. entrepreneur Ryan Hoover. It was admitted to Y Combinator in 2014, and received funds from several prominent venture capitalists, including Greylock, and was finally acquired in 2016 by AngelList.

The products that are posted on Product Hunt can be web apps, mobile apps, hardware products, games, and even books and podcasts. Users submit products, which are then listed on the website. With thousands of products listed each year, Product Hunt has become the website where startups launch their products to gain visibility and attract funds.[9] We retrieved the entire dataset of products featured on Product Hunt from January 2014 to September 2022. For this period, we were able to match startups with their product launches, using the startups' website information. As shown in Table 1a, 6% of the startups founded after 2011 had launched a product on Product Hunt

[8] As a robustness check, 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. We exclude from the categorization approximately 18% of the total repositories for which we could not identify a coherent category.

[9] `https://www.reddit.com/r/SideProject/comments/qhpjau/how_to_launch_on_product_hunt_a_detailed_guide/`.

as of December 2022.

4 Empirical specification

Ideally, to assess how engaging with an open source community is related to attracting funds, we would randomly assign startups to the treatment of interest. That way, we would address the concern that the relationship between a startup's engagement with OSCs on GitHub and achieving funding milestones may be confounded by technology or geographical shocks, a startup's technology characteristics, or a startup's technology lifecycle. Implementing a field experiment that randomizes access to OSCs across startups and over time would be very costly and could raise important ethical concerns as discussed in the Belmont Report.[10] In its place, we estimate a difference-in-differences model where we compare startups that engage with OSCs on GitHub to randomly selected control units among a subgroup of startups with similar observable characteristics. By doing so, we assess how the likelihood of being funded by a given period changes after treated startups begin to engage with OSCs on GitHub relative to the control group. To build the control group, we randomly select up to five startups that are either inactive on GitHub or only engage in activities related to their internal repositories (this occurrence is more rare) and were founded in the same state and year as the treated startup, and have a similar value of the software share index described above.[11] We additionally impose that a treated startup and its control group have similar human capital, as defined by whether a startup's founder or CXO[12] is highly ranked on Crunchbase's list of top people.[13] By doing so, we compare, as much as possible, treated startups with controls exposed to similar funding conditions and technology shocks, developing similar technologies, operating in a similar phase of the technology life cycle, and characterized by comparable human capital. Consequently, for every startup i that engaged in at least one external activity on GitHub, and associated with control startups in group g, we estimate:

$$Y_{igt} = \delta PostGitHub_{gt} \times EngagementWithOSC_i + \mu_i + \psi_{gt} + \tau_{it} + \varepsilon_{igt}, \tag{1}$$

[10] `https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html`

[11] Specifically, we constructed ten bins for the software share index with cutoff values at $0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$ and derived control startups among those whose index belongs to the same software share bin as the treated startup.

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

[13] Refer to: `https://www.crunchbase.com/discover/people`, accessed March 4, 2022. We consider the first 1,000 entrepreneurs on Crunchbase's list as top-ranked.

where Y_{igt} is an indicator that becomes one starting from the quarter t when startup i, belonging to the treated-control group g, raises a first financing round. We focus on a startup's first round as this is a fundamental funding milestone. Indeed, investors participating in this round typically provide the necessary network for raising subsequent rounds (Conti and Graham, 2020). We examine variants of this outcome, delving into the types of investors a startup attracts in the first financing round, and the amount raised. We additionally examine whether a startup launched a product listed on Product Hunt. We conduct the analysis at the quarter level to mitigate the concern that even though we have monthly data on the round announcements, startups are informed by their investors a few months before the round announcement date. We censor the sample in the second quarter of 2020, the date at which we retrieved the Crunchbase dataset. Moreover, following Conti and Guzman (2021) and Amore et al. (2023), we observe each startup for a maximum of six years, as most startups raising a financing round do so within this time window.

$EngagementWithOSC_i$ is our treatment indicator, which takes a value of one if startup i engages in at least one public activity with external repositories on GitHub during the period we observe. In more fine-grained analyses, we also distinguish a startup's type of engagement with OSCs on GitHub. $PostGitHub_{gt}$ is a time-varying binary indicator that becomes one for all the startups in a treated-control group g after startup i engages in an external activity on GitHub. The coefficient of interest is δ, which is associated with the interaction between $EngagementWithOSC_i$ and $PostGitHub_{gt}$. This coefficient measures the average change in the likelihood that a startup has raised a first financing round by quarter t after it engages with OSCs on GitHub. In our specifications, the effects of $EngagementWithOSC_i$ and $PostGitHub_{gt}$ are absorbed by our fixed effects.

While our empirical approach enables comparison of observably similar startups, it may still suffer from omitted variable bias to the extent that our selection of the control group based on observables does not allow us to fully net out the impact of a startup's intrinsic characteristics, as well as technology trends. To address this concern, we saturate our main specification with relevant fixed effects. Specifically, μ_i is a fixed effect for startup i, which absorbs differences in startups' fixed characteristics. The ψ_{gt} is a group g by (year-)quarter fixed effect, which controls for the possibility that differences across treated-control groups may change over time. Moreover, τ_{it} is a startup's industry group by quarter fixed effect that absorbs the effect of technology shocks we might have been unable to capture with our selection of the control group. We consider the same industry

groups as those utilized to build the synthetic software share index. We also add startup by age fixed effects, where age is defined at the yearly level to provide an even more stringent control of startup technology life cycle effects.

Descriptive statistics for the matched sample are reported in Table 2a. In Table 2b, we further distinguish between startups that engaged with with OSCs on GitHub and the control group. Relative to the control group, the proportion of treated startups that have raised a financing round, and have raised a first round from a VC and/or a successful investor is larger. Additionally, treated startups raise larger rounds than untreated startups. Finally, the proportion of startups that have launched a product on Product Hunt is larger among treated companies than among the control group.

⟨ Insert Table 2 about here ⟩

5 Results

5.1 Open source communities on GitHub and attracting funds: Baseline results

We begin by discussing the results of our difference-in-differences model, where we compare the change in the probability that a startup will have raised a financing round after it starts to engage with OSCs on GitHub, relative to a control group of startups with no such engagement. The results are reported in Table 3, where we cluster standard errors by groups of treated startups and their controls.[14]

Our approach consists of progressively saturating our model with fixed effects to address potential threats from factors we might not have appropriately controlled for. The model in column 1 includes treated-control group, year-quarter, and startup fixed effects. The latter control for time-invariant unobservables across startups that may be correlated with the propensity to engage with OSCs. As shown, post-GitHub involvement, startups are more likely to have obtained funds, suggesting that they become more attractive to potential investors. Specifically, startups that become actively engaged with external repositories available on GitHub are 15 percentage points more likely to have raised a financing round relative to their controls (p-value=0.00), all else equal. This represents a 65% increase in the outcome mean of 0.23. The results remain similar in column 2, where we add treated-control group by year-quarter fixed effects, which address the possibility that differences

[14]Standard errors do not substantially change if we cluster them at the startup level. However, we prefer to cluster standard errors by groups of treated startups and their controls because this method tends to produce more conservative (that is, larger) confidence intervals.

between startups that engage with OSCs and their controls may change over time. In column 3, we add industry group by year fixed effects. These fixed effects address the possibility that technology trends differentially impact startups' ability to attract funds. With this specification, the magnitude of the coefficient of interest slightly decreases to 12 percentage points. Finally, in column 4, the main effect becomes 0.08 – equivalent to a 36% increase in the mean –, once we add age (measured in years) by startup fixed effects to account for differences in firms' technology life cycle which may systematically vary with their engagement with OSCs on GitHub and be correlated with the outcome. Taken together, these results suggest that participation in OSCs can help startups achieve funding milestones, possibly by accelerating technology development. As shown in Table A1, these results are robust to not matching treated and control startups on *Top Team*, which may not be a strictly predetermined control. Additionally, Table A2 shows that the results are robust to a modified matching procedure where we required treated and control startups to share at least one technology keyword rather than a similar value of the *software share* index.

In Figure 2, we explore possible differences in pre-trends between treated and control startups. For this purpose, we modify Eq. (1), substituting the $PostGitHub_{gt}$ indicator with dummies for each quarter before and after a startup becomes engaged with OSCs on GitHub. We restrict the sample to the four quarters preceding a startup's involvement on GitHub and the eight quarters after, to focus on the period just before and after such an involvement. We control for the same set of fixed effects as in column 1 of Table 3.

We infer two important implications from this figure. First, the difference in the probability of having obtained funds is flat around zero in the pre-period, suggesting that startups becoming involved with OSCs on GitHub and their controls do not differ in meaningful ways, conditional on our set of fixed effects.[15] Second, the relationship between a startup's engagement with OSCs and the likelihood of being funded is strong, immediate, and persists over time.[16]

⟨ Insert Table 3 and Figure 2 about here ⟩

[15] In Figure A3 of the Appendix, we provide further robustness tests suggesting that the parallel trends assumption holds within large bounds, using the approach of Rambachan and Roth (2023).

[16] The results are also robust to adopting a staggered difference-in-differences approach with time and firm fixed effects using Callaway and Sant'Anna (2021)'s method. In applying their method, we do not need to rely on a matched control group and can compare the probability of receiving funding before and after becoming active on GitHub within the sample of treated firms. The estimate of the ATT (simple average) is 0.058 with a standard error of 0.001.

5.2 The life cycle of a startup's technology and engagement with open source communities on GitHub

While the relationships we report in Table 3 are robust to different fixed-effect specifications, we might still omit some factors that could bias our estimates. For instance, it is possible that the timing when a startup starts to engage with OSCs on GitHub is correlated with the startup's technology life cycle. The inclusion of startup by age fixed effects should at least in part control for such concerns, but age is measured in years rather than in months. To go a step further, we use information on a startup's activities related to its own *internal* repositories on GitHub. We thus modify Eq. (1), this time considering two treatments: (1) whether a startup engages with OSCs on GitHub by interacting with repositories it does not directly control, and (2) whether a startup uses GitHub as a public software development tool, engaging with internal repositories it directly controls. The reference group is represented by startups that are randomly selected with the same criteria mentioned above, and are completely inactive on GitHub. The logic is that engagement with external and internal repositories should be subject to similar technology life cycle effects. Therefore, controlling for a startup's engagement with internal repositories should allow us to net out life cycle effects more precisely. The results from this exercise are reported in Table 4. As shown, the association between startups that become engaged with OSCs on GitHub through interactions with external repositories and receiving funding is nine percentage points larger than the association between startups engaging in activities related to their own internal repositories and receiving funding, all relative to the reference group (p-value=0.00). These results provide a strong indication that technology life cycle effects may not be driving our results.

⟨ Insert Table 4 about here ⟩

5.3 Opportunity costs of engaging with open source communities on GitHub

In an alternative exercise, we first assess which startups have lower opportunity costs to become engaged with OSCs. Using these startups, we then compare the relationship of their engagement with OSCs on GitHub with receiving funding to the relationship of those startups with higher opportunity costs of engagement on GitHub. The rationale is to approximate the control group of startups unengaged with OSCs -– who arguably have such high opportunity costs as to preclude them from engaging with GitHub – with engaged startups that are on the high end of opportunity

costs associated with interacting with OSCs. We include this analysis to further mitigate omitted variable bias concerns.

Given our discussion in Section 2, we consider the following two determinants of a startup's opportunity costs of becoming involved with OSCs on GitHub: the novelty of a startup's technology and the level of market competition. On the one hand, a startup's benefits from interacting with OSCs may be substantial when a startup's technology is novel, as its development may require solving relatively complex problems. Further, when startups develop novel technologies, they might have an incentive to rapidly push them to the market to become market leaders, and, to achieve this, they may rely on relatively cheap and readily-available open source input. On the other hand, appropriability issues may be more severe when there are many other startups operating in the technology space of the focal company. In this exercise, product and market features act as drivers of the relationship between startups' interactions with GitHub and their likelihood of being funded. We measure novelty by whether the combination of a startup's industry group keywords at the startup's founding year is relatively new.[17] That is, less than three years should have passed since a keyword combination first appeared on Crunchbase.[18] To operationalize the level of competition in each technology space, we count the number of startups active in a given year-quarter and with the same industry group keywords as the focal one. The market in which a focal startup operates is thus considered competitive if the number of startups possessing the same technology keyword combination is greater than the 90^{th} percentile.

We first estimate a linear probability model where the dependent variable is $EngagementWithOSC_i$. As reported in Figure A4 of the Appendix, startups developing novel technologies are nine percentage points more likely to engage with OSCs on GitHub. Conversely, startups that have experienced high competition for at least one year-quarter are two percentage points less likely to engage with OSCs on GitHub.

Moving to Table 5, the results in column 1 show that startups developing relatively new technologies become eight percentage points more likely to have attracted funds by a given quarter after engaging with OSCs on GitHub relative to other startups that engage with OSCs. These results are obtained after accounting for the differential propensity of novel technologies to receive funding in any given

[17] We observe 7,516 unique combinations of industry group keywords in our dataset.

[18] Using alternative cutoffs does not change the results.

period, which we hold constant by including year-quarter by technology novelty fixed effects. We include these fixed effects in addition to startup and year-quarter by treated-control group fixed effects.

Moreover, as reported in column 2 of Table 5, startups operating in competitive markets are 14 percentage points less likely to have been financed after they become engaged with OSCs, relative to other startups that engage with OSCs. Computing the linear combination of $PostGitHub_{gt} \times EngagementWithOSC_i$ and $HighCompetition_{it} \times PostGitHub_{gt} \times EngagementWithOSC_i$ reveals that the overall impact of engaging with OSCs on GitHub on the likelihood of being financed is zero for startups operating in competitive markets relative to the reference outcome. Here, again, we account for the differential propensity of startups in competitive markets to receive funding by including year-quarter by competitive market fixed effects. Overall, these results suggest that engaging with OSCs is especially beneficial when the opportunity costs of doing so are relatively low.

⟨ Insert Table 5 about here ⟩

5.4 Exogenous variation in the costs of engaging with open source communities on GitHub

To corroborate our findings so far, we further exploit exogenous variation in the opportunity costs of interacting with OSCs on GitHub's code hosting website, *github.com*. GitHub offers multiple products, of which two are important in our context. One is *github.com*, which we have discussed so far. The second product is *GitHub Enterprise (GHE)*, which has essentially the same functionality as *github.com*, but it is not designed to be public. First launched in November 2011, GHE is a self-hosted platform for software development *within* an enterprise.[19] Employees of an enterprise can use GHE to build and share software using Git version control as they would do on *github.com*. However, GHE runs on an enterprise's infrastructure and is governed by access and security controls the enterprise defines. As pointed out in several outlets, the benefits of using GHE are considerable and can be summarized into enhanced security, better collaboration, increased scalability, and automated workflows.[20]

[19] See `https://docs.github.com/en/enterprise-server@3.5/admin/overview/about-github-enterprise-server`.

[20] See, for example, https://www.gitkraken.com/blog/top-5-reasons-to-use-github-enterprise and https://github.com/enterprise.

While GHE is appealing to any firm, startups, which typically lack a server infrastructure, only began to consistently use it when made available as a virtual private cloud. Specifically, after November 2014, GitHub enabled organizations to run their own GitHub Enterprise server through Amazon Web Services (AWS).[21] Relevant to our analysis is the fact that the integration of GHE with AWS must have lowered the costs of developing proprietary software relative to engaging with OSCs, for all startups, regardless of the life cycle, and other time-varying characteristics, of their technology. Therefore, the exploitation of such a shock can potentially set us closer to addressing concerns of omitted variable bias.

As shown in Figure A5, the startups in our dataset sharply decreased their engagement with OSCs on *github.com* after GHE became available on AWS. This pattern is likely driven by startups migrating their software development efforts to the virtual private cloud, as the opportunity costs of interacting with repositories of others might have increased.

We exploit the launch of GHE on AWS in an analysis restricted to the subset of startups active on *github.com*. Specifically, we assess how the likelihood of being funded changes for startups that engage with OSCs on GitHub in a given quarter before GHE becomes available on AWS relative to after. We restrict the analysis to the quarters just before and after the integration of GHE with AWS to limit the concern that other macroeconomic trends occurring around the shock could confound the association of interest.

In practice, we estimate a modified version of Eq. (1) for the subsample of treated startups, interacting a startup's number of engagements with OSCs in quarter t with $PostAWS_t$, an indicator identifying the period after the last quarter of 2014 when GHE could be installed on AWS.[22] We include year-quarter by industry group fixed effects to absorb the effects of macroeconomic shocks that may vary by technology. Moreover, we augment the equation with startup fixed effects to at least partially address the concern that while the timing of the launch of GHE on AWS is likely unrelated to factors such as time-invariant startup characteristics, the actual adoption of GHE on AWS may be non-random and startups selected into using GHE on AWS for unobserved reasons.

The results are reported in Table 6 and Figure 3. As we limit the analysis to the subsample of

[21] See `https://github.blog/2014-11-11-a-faster-more-flexible-github-enterprise/`. AWS is very popular among startups, which have benefited from various discounts AWS has offered over the years. See `https://aws.amazon.com/about-aws/whats-new/2013/10/10/announcing-aws-activate-a-new-global-program-for-startups/`.

[22] We replace the indicator for whether a startup becomes engaged with OSCs with the number of engagements to compensate for the decrease in variation over time provided we examine a shorter time span.

treated startups, we cluster standard errors at the startup level. As shown, the relationship between the number of engagements with OSCs and the likelihood of being funded is stronger after GHE becomes available on AWS than before. Reassuringly, the quarterly differences in pre-trends in Figure 3 tend all to be close to zero.

⟨ Insert Table 6 and Figure 3 about here ⟩

Given that, after GHE was enabled on AWS, the opportunity costs of engaging with OSCs increased, we interpret the results in Table 6 such that startups interact with OSCs through *github.com*, only when the value of doing so is large enough to offset the associated costs, with a resulting positive impact on achieving funding milestones. To bolster this interpretation, we modify the model in Table 6, assessing how the impact of the integration of GHE with AWS interacts with the novelty of a startup's technology and the level of competition. The rationale is that, after GHE becomes available on AWS, the incentives to engage with OSCs on GitHub likely differ for these startup types relative to other startups. Indeed, in unreported results, we find that startups developing novel technologies are 41% more likely than other startups to engage with OSCs on GitHub after GHE became available on AWS (p-value=0.00). Conversely, startups that operate in highly competitive markets are 31% less likely to do so (p-value=0.00). However, these differences should not be correlated with a startup's specific phase in its technology life cycle.

As shown in column 1 of Table 7, the intensity of the engagement of startups developing novel technologies with OSCs on GitHub does not seem to make these startups more successful at attracting investors relative to other startups before GHE becomes available on AWS. However, after GHE becomes available on AWS, we observe that the effect of the number of engagements with OSCs on the likelihood of being funded is stronger (p-value=0.00) for startups developing novel technologies relative to the other startups. These results are obtained by including year-quarter by technology novelty fixed effects, in addition to startup and year-quarter by treated-control group fixed effects.

In column 2 of Table 7, we distinguish startups by how competitive the market in which they operate is. Here, startups operating in competitive markets display the weakest relationship between the number of engagements with OSCs after GHE becomes available on AWS and the likelihood of being funded (p-value=0.04). This difference in effects is 34 times larger than the difference observed before the integration of GHE with AWS (which is only significant with a p-value of 0.061).

Overall, these exercises suggest that startups with the lowest opportunity costs of interacting with OSCs on GitHub enjoyed the largest gains from such interactions after the integration of GHE with AWS unlocked new opportunities for within-firm collaborations.

⟨ Insert Table 7 about here ⟩

5.5 Startup technologies and types of external repositories they engage with

To enrich our analysis, we categorize the technology use-cases of external repositories with which startups interact. In particular, we distinguish between engagement with OSCs with respect to SD/BE, ML, API, and UI. Further, we group a subsample of observed startups into different domains: Software Tools, AI & Blockchain, Platform, Consumer-facing, and the remainder. The goal of this exercise is twofold. First, we want to examine the association between startup domains and the technology use cases of external repositories they interact with on GitHub. Second, we plan to discern which startups gain the most by interacting with (what type of) external repositories on GitHub.

To achieve the first goal, we produce cross-sectional correlations in Figure 4 between the use-cases of the external repositories GitHub-active startups interact with and the domains in which these startups specialize, having controlled for treated-control group and startup founding year fixed effects and clustered standard errors by treated-control group. We standardize the coefficients with the means of the corresponding outcomes. As shown in Figure 4, platform startups are the most active with external repositories on GitHub across all use cases. This is expected given that their business model rests on cooperating with developers outside a startup (Benzell et al., 2019), and therefore may have a better product-market fit. Platform startups are followed by Software Tools and AI & Blockchain startups. The latter category especially engages with external ML repositories. Finally, Consumer-facing startups are the least active on the GitHub platform.

⟨ Insert Figure 4 about here ⟩

We next assess which startups gain the most from engaging with OSCs on GitHub. We do so by interacting the different terms on the right-hand side of Eq. (1) with the startup domain indicators generated. We add treated-control group and startup fixed effects, as well as year-quarter by startup domain fixed effects. The domains we consider are Software Tools, AI & Blockchain, Platform, and Consumer-facing.

The results reported in Figure 5 show that AI & Blockchain startups derive the largest benefits, in terms of improved likelihood of attracting funds, from interacting with OSCs on GitHub. As AI & Blockchain are relatively more novel and complex technologies, these results are consistent with our earlier findings that novel technologies derive the largest financing gains from the interaction with OSCs on GitHub.

We delve deeper into this last result in Figure 6, where we distinguish the different technology use-cases with which our startups engage on GitHub. In Panel A, we find that startups derive positive gains by interacting with the full spectrum of repository use cases. Zooming in on AI & Blockchain startups in Panel B, we show that they do not derive as much gains by engaging with external repositories related to UI.

⟨ Insert Figure 5 and Figure 6 about here ⟩

5.6 Engagement with open source for visibility or technology development?

Having shown that engaging with OSCs on GitHub is related to attracting funds, we next evaluate whether startups become active on GitHub to merely enhance their visibility or to also improve their technology. In Table 8, we examine the impact of an activity that startups should pursue if their main objective is to gain visibility vis-à-vis potential investors. We focus on the creation or modification of their repositories' readme files, which provide documentation to a repository that can include a non-technical summary of the codebase in the repository. We estimate a similar model as reported in Table 5, decomposing internal activities into creation or modification of readme files and the remainder. As shown, creating or modifying readme files is not significantly related to receiving funding (p-value=0.76). As such, this result corroborates our earlier findings, perhaps indicating that investors do not simply react to "cosmetic" activities startups do on GitHub.

⟨ Insert Table 8 about here ⟩

To bring our results full circle, we assess whether a startup's engagement with OSCs on GitHub is related to the creation of technology products that potential investors may value. For this scope, we use Product Hunt data on product launches. As Product Hunt only started in 2014, we restrict our sample to those startups founded after 2011 for this analysis, given that older startups might have already launched their products, but without relying on the Product Hunt platform.[23] Building on these data, we modify Eq. (1), considering as an outcome whether a startup will have launched a

[23] Using more stringent cutoffs, such as considering startups that were founded after 2013 does not change the results.

product by quarter t. The results are reported in Table 9, reproducing the same structure as in Table 3. For example, in columns 1 to 3, we show that after startups begin to engage with OSCs on GitHub, their likelihood to have launched a product increases by eight percentage points, which represents an increase by a factor of 1.8 in the outcome mean. In Table A3, we show that the results remain similar when we only consider higher-quality products, that is, those whose number of ratings or upvotes received as of September 2022 is greater than the 75 percentile. Moreover, the results in Table A4 show that the association is stronger when we examine the likelihood that a startup will have launched the first version of a product. This last result is in line with our findings on technology novelty and suggests that the gains from interacting with OSCs are more substantial when startups can address the market with novel products.

Overall, the results using data from Crunchbase as well as Product Hunt suggest that startups rely on knowledge generated by engaging with OSCs to produce palpable products that are valued by potential investors. Put differently, our findings provide suggestive evidence that startups are, indeed, "beefing up" their technology to receive early-stage financing by relying on OSCs rather than *merely* increasing their visibility to potential investors.

⟨ Insert Table 9 about here ⟩

5.7 For which types of investors are the results strongest?

In the last part of our empirical analysis, we aim to assess which investors are most responsive to startups that engage with OSCs on GitHub. The analyses are reported in Table 10. We begin by considering whether a startup's engagement with OSCs is related to raising large round amounts (column 1) or smaller ones (column 2). Large financing amounts are those in the upper quartile for first rounds. As shown, after startups engage with OSCs, their likelihood of being funded through a large first round increases by seven percentage points, equivalent to an increase of a factor of 1.4 from the mean. Conversely, their likelihood of being funded through a smaller round increases by five percentage points, relative to a mean of 0.18.

Further, we show in columns 3 and 4 that the relationship between startups engaging with OSCs and the likelihood that they attract VC funds is larger by a factor of 1.5 relative to the mean, while their likelihood of attracting non-VC investors increases by 1.6 percentage points relative to a mean of 0.15. Finally, we find similar differences in the last two columns of Table 10, where we distinguish between having raised a first round from successful investors (column 5) and having raised a first

round from less successful investors (column 6). We measure investor success by the number of portfolio startups that were acquired or went public in the five years prior to investing in the focal startup. Successful investors are those with a number of portfolio exits in the upper quartile.

Overall, these results suggest that by engaging with OSCs, startups can raise relatively larger financing rounds from investors that provide large non-financial capital in addition to money.

⟨ Insert Table 10 about here ⟩

6 Discussion and Conclusions

In this paper, we investigate the role of engaging with OSCs among nascent firms in raising funds. We analyze unique data linking Crunchbase profiles to accounts on the open source software development platform GitHub, and to product launches on the Product Hunt website. Our results suggest that participation in OSCs can play an important role in achieving funding milestones. Applying a difference-in-differences approach to estimate the likelihood of being funded as a function of engaging with OSCs on GitHub, we find that such engagement substantially increases the likelihood that a startup will have raised a first financing round. The relationship is strong. In fact, the estimated coefficients indicate at least a 36% increase in the mean. These findings are corroborated exploiting a quasi-experimental change in the relative cost of engaging with OSCs, and appear to be most pronounced when startups raise large rounds and when funds are obtained from VCs and successful investors. These results, taken together, provide an indication that a startup's involvement with open source communities may be crucial for attracting earliest-round valuable investors.

Moreover, our results further reveal that the relationship is strongest when startups are developing novel technologies, for which the benefits of relying on knowledge inputs from the open source (e.g., speed, access to existing integration packages, co-production to make the technology work) may offset costs (Jordan and Mitchell, 2015). Conversely, the relationship is weakest when the level of competition is high, and thus costs, such as appropriability concerns, are potentially more relevant. These findings also translate when we more closely examine the precise domain a startup is active in. Using information from Crunchbase to group firms into AI & Blockchain, Consumer-Facing, Platform, and Software Tools our results suggest that startups active in AI & Blockchain — the on average most novel domain -– are those who achieve the highest gain. These firms typically rely on input by others to create a functioning product (e.g., state-of-the art ML algorithms). Conversely,

firms operating in the Consumer-Facing domain experience the lowest benefits from engaging with OSCs.

We provide evidence suggesting that startups do not merely engage with OSCs to increase their visibility vis-a-vis investors. It is possible that startups rely on OSCs and a public platform such as Github 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). However, our findings can also be interpreted such that startups may be using code repositories available on GitHub to engage with external knowledge critical for scaling, integrating into the ecosystem, and producing a, at least, minimal viable product. Consistent with that interpretation, we show that investors do not react to "cosmetic" changes startups add to the readme files associated with their repositories, which would primarily enhance the startups' visibility. Finally, we find that engagement on GitHub is associated with the launching of new technology products, especially high-quality ones, providing suggestive evidence that open source inputs allow startups to upgrade their technology stack.

Although we cannot exclude that GitHub participation also increases the visibility of a new venture, our interpretation of the results is that increasing visibility is unlikely the sole purpose of engagement with OSCs on GitHub. On the whole, our findings seem to indicate that startups are feasibly "beefing up" their technologies by relying on inputs from OSCs, and by doing so, they increase their likelihood of receiving early-stage financing.

Our study contributes to increasing our understanding of the role of a particular channel through which startups can access outside knowledge – open source –, and how open source matters in attracting funding. As such, our findings extend the literature that analyzes 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 with OSCs. Namely, in our context, technology startups rely on open source to attract 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 an advantage of open source engagement lies in development, scale, and integration, 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.

In conclusion, this paper provides important insight into the role of engaging with external knowledge through open source platforms for startup performance outcomes. In particular, our findings contribute to our understanding of the impact of early-stage tech-stack investments on achieving funding milestones (Roche et al., 2020a). By opening the technology "black-box", we reveal important nuances that have been largely overlooked in the literature, namely that using open source to build an, at least minimum viable product, can help firms attract funding. 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.

Almirall, E., and Casadesus-Masanell, R. (2010). "Open versus closed innovation: A model of discovery and divergence." *Academy of Management Review*, *35*(1), 27–47.

Amore, M. D., Conti, A., and Pelucco, V. (2023). "Micro venture capital." *Strategic Entrepreneurship Journal*.

Benzell, S., Hersh, J. S., Van Alstyne, M. W., and Lagarda, G. (2019). "How apis create growth by inverting the firm." *Available at SSRN 3432591*.

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.

Callaway, B., and Sant'Anna, P. H. (2021). "Difference-in-differences with multiple time periods." *Journal of Econometrics*, *225*(2), 200–230.

Chesbrough, H. (2006). "New puzzles and new findings." In H. Chesbrough, W. Vanhaverbeke, and I. West (Eds.), *Open Innovation: Researching a New Paradigm*, Oxford, UK: Oxford University Press.

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., and Graham, S. J. (2020). "Valuable choices: prominent venture capitalists' influence on startup ceo replacements." *Management Science*, *66*(3), 1325–1350.

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

Conti, A., and Guzman, J. A. (2021). "What Is the US Comparative Advantage in Entrepreneurship? Evidence from Israeli Migration to the United States." *The Review of Economics and Statistics*, 1–45.

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.

Dahlander, L., and Magnusson, M. (2008). "How do firms make use of open source communities?" *Long Range Planning*, *41*(6), 629–649.

Dahlander, L., O'Mahony, S., and Gann, D. M. (2016). "One foot in, one foot out: how does individuals' external search breadth affect innovation outcomes?" *Strategic Management Journal*, *37*(2), 280–302.

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.

Fleming, L. (2001). "Recombinant uncertainty in technological search." *Management Science*, *47*(1), 117–132.

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.

Grimpe, C., and Kaiser, U. (2010). "Balancing internal and external knowledge acquisition: the gains and pains from R&D outsourcing." *Journal of Management Studies*, *47*(8), 1483–1509.

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

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.

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

Jordan, M. I., and Mitchell, T. M. (2015). "Machine learning: Trends, perspectives, and prospects." *Science*, *349*(6245), 255–260.

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.

Laursen, K., and Salter, A. (2006). "Open for innovation: the role of openness in explaining innovation performance among uk manufacturing firms." *Strategic Management Journal*, *27*(2), 131–150.

Laursen, K., and Salter, A. J. (2014). "The paradox of openness: Appropriability, external search and collaboration." *Research Policy*, *43*(5), 867–878.

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*.

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.

Rambachan, A., and Roth, J. (2023). "A more credible approach to parallel trends." *Review of Economic Studies*, rdad018.

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

Roche, M. P., Conti, A., and Rothaermel, F. T. (2020b). "Different founders, different venture outcomes: A comparative analysis of academic and non-academic startups." *Research Policy*, *49*(10), 104062.

Rysman, M., and Simcoe, T. (2008). "Patents and the performance of voluntary standard-setting organizations." *Management Science*, *54*(11), 1920–1934.

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.

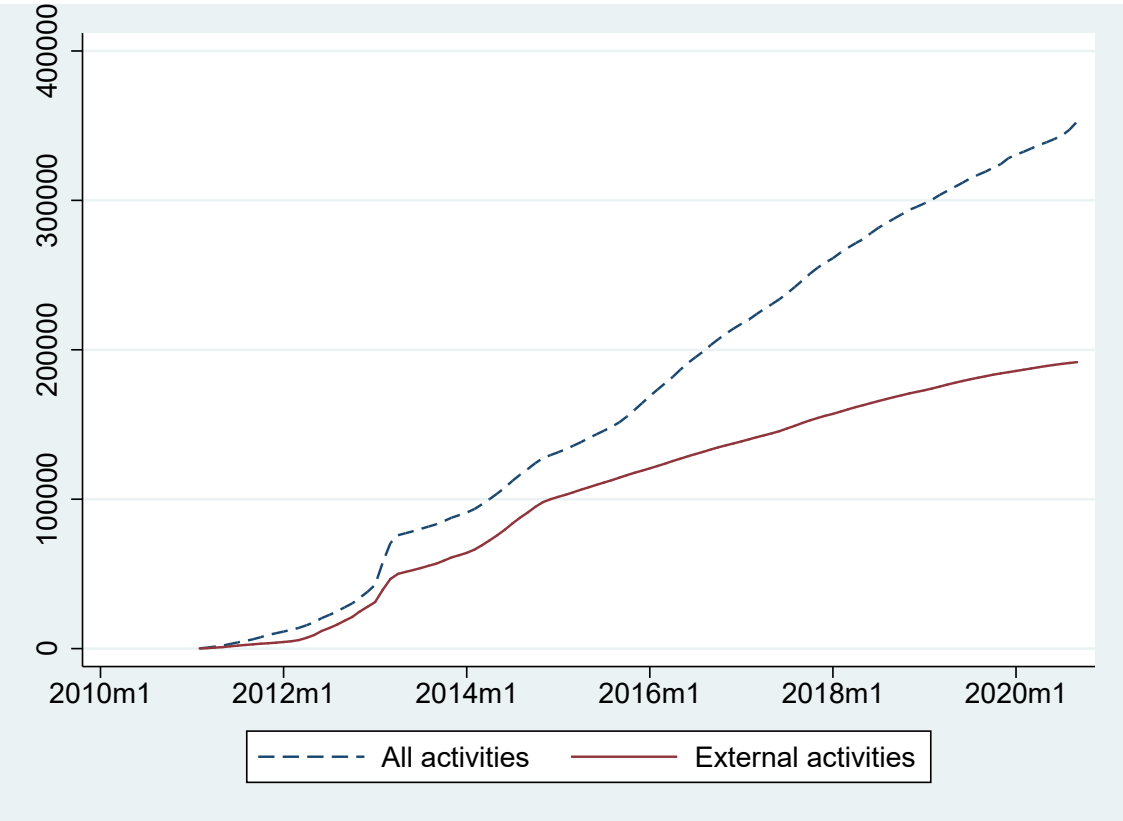
Spence, M. (2002). "Signaling in retrospect and the informational structure of markets." *American Economic Review*, *92*(3), 434–459.

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

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

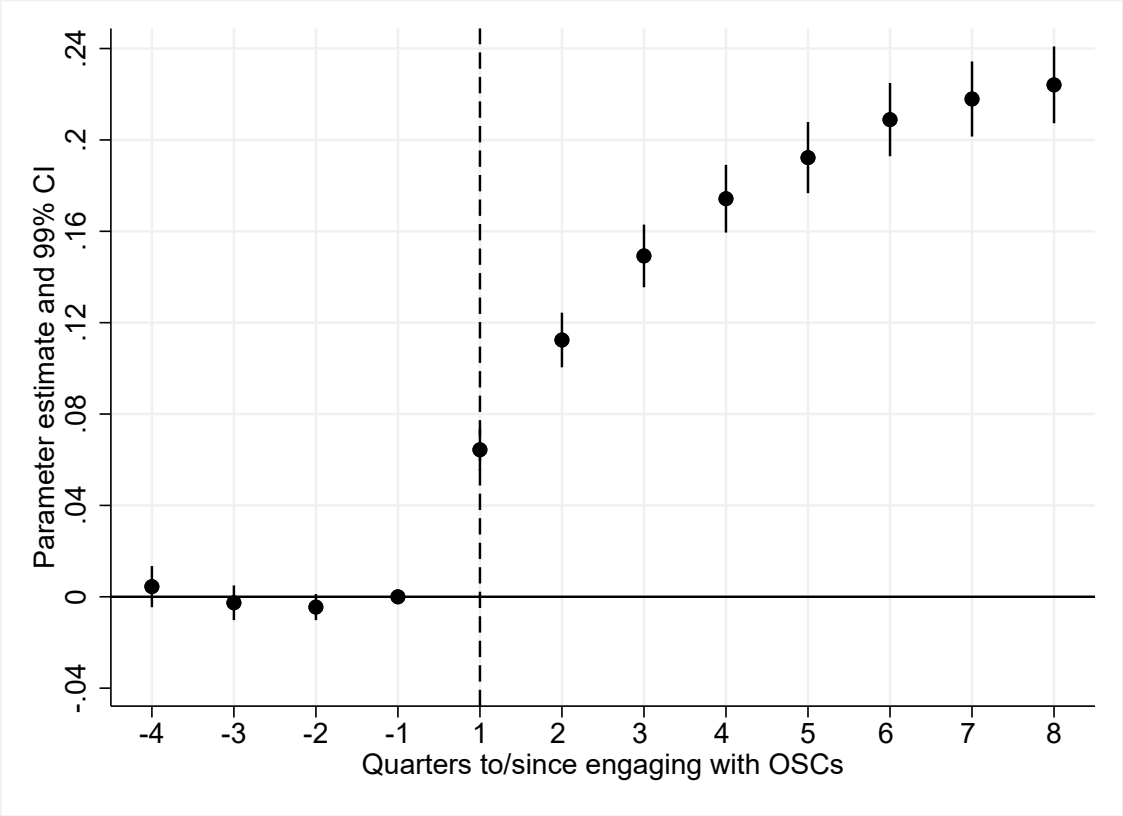
Yoo, Y., Boland Jr, R. J., Lyytinen, K., and Majchrzak, A. (2012). "Organizing for innovation in the digitized world." *Organization Science*, *23*(5), 1398–1408.

Figure 1: Startups' engagements with external and total GitHub repositories over time



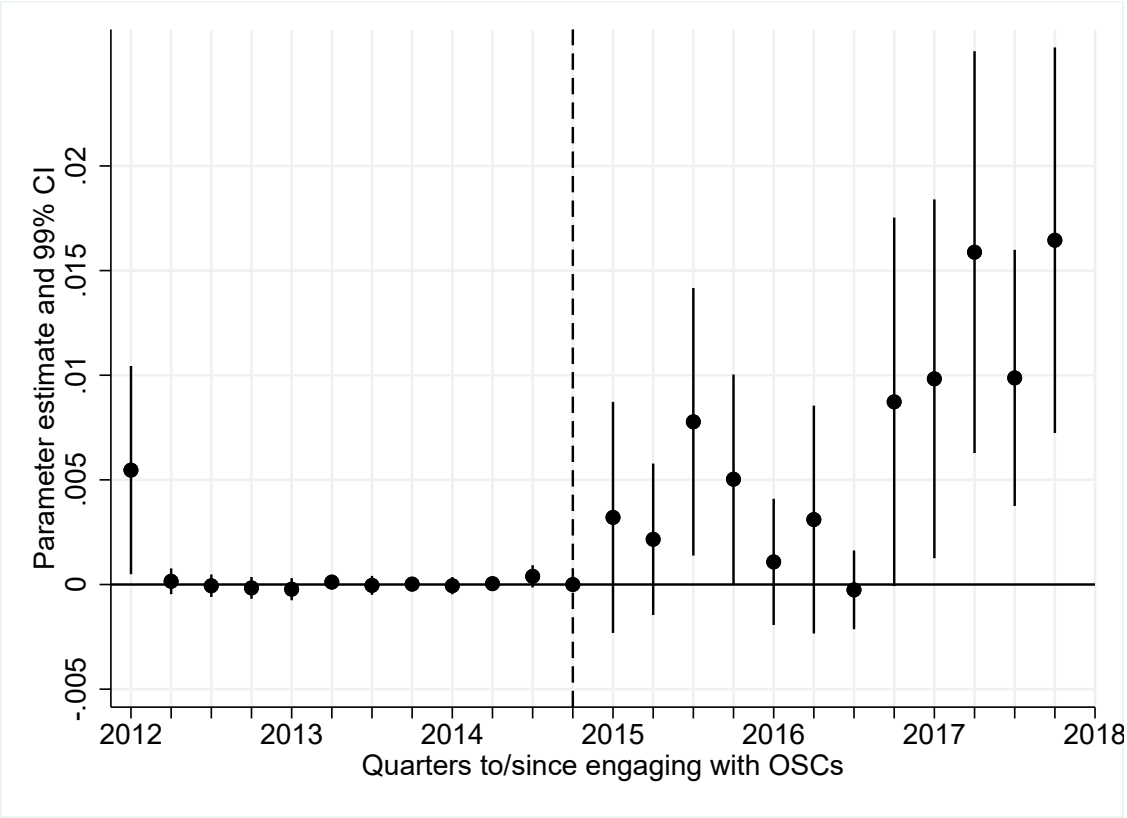
Notes: This figure displays patterns of activity on GitHub over time. We distinguish between external, and all activities (which include external activities).

Figure 2: Raising a first financing round: Event study



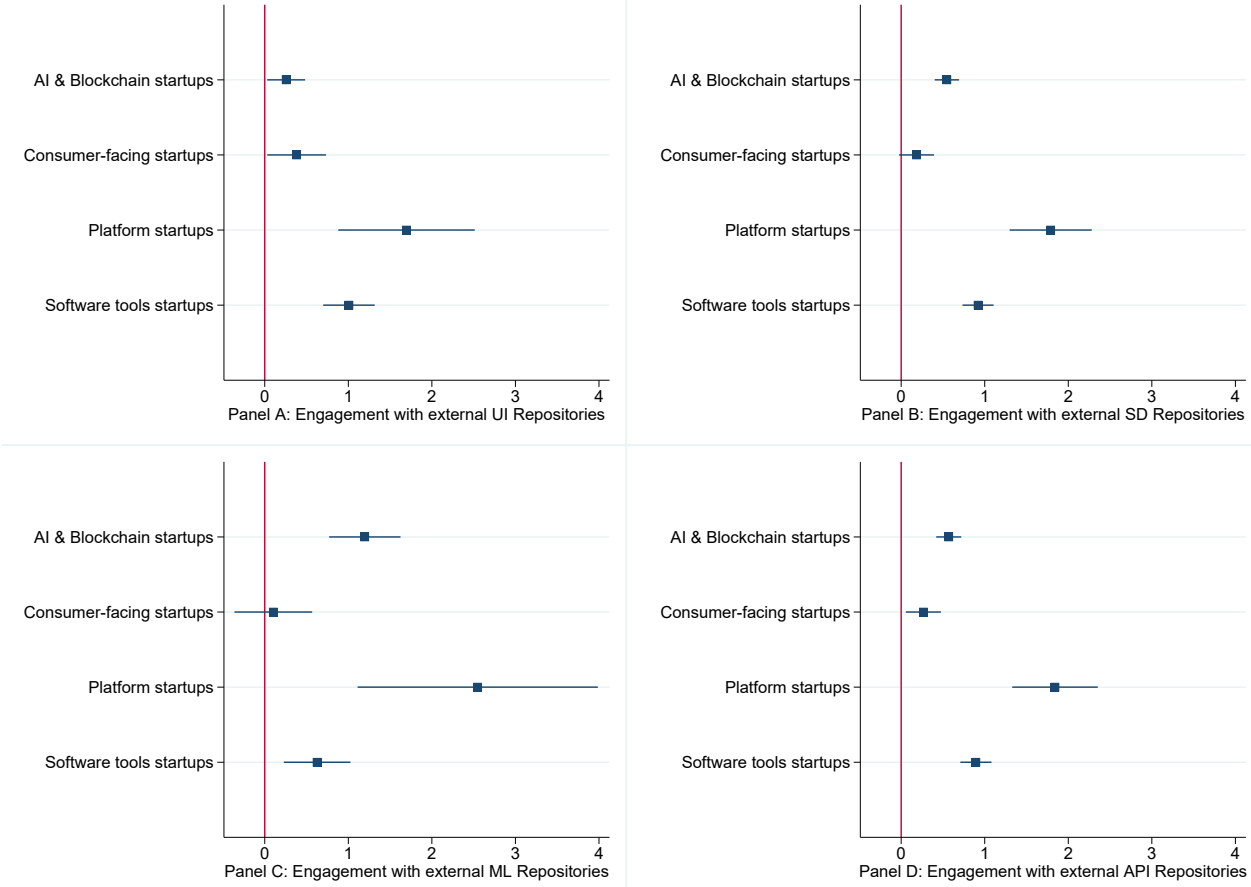
Notes: This figure shows how the probability that a startup will have raised a round changes after a startup starts engaging with OSCs on GitHub. To generate this graph, we modified Eq. (1) in the main text by substituting the $PostGitHub_{g,t}$ indicator with binary variables for each of the pre- and post-treatment years. We control for the same fixed effects as those in column 1 of Table 3. The vertical lines represent 99% confidence intervals.

Figure 3: Raising a first financing round: Event study applied to startups that had engaged with OSCs on GitHub



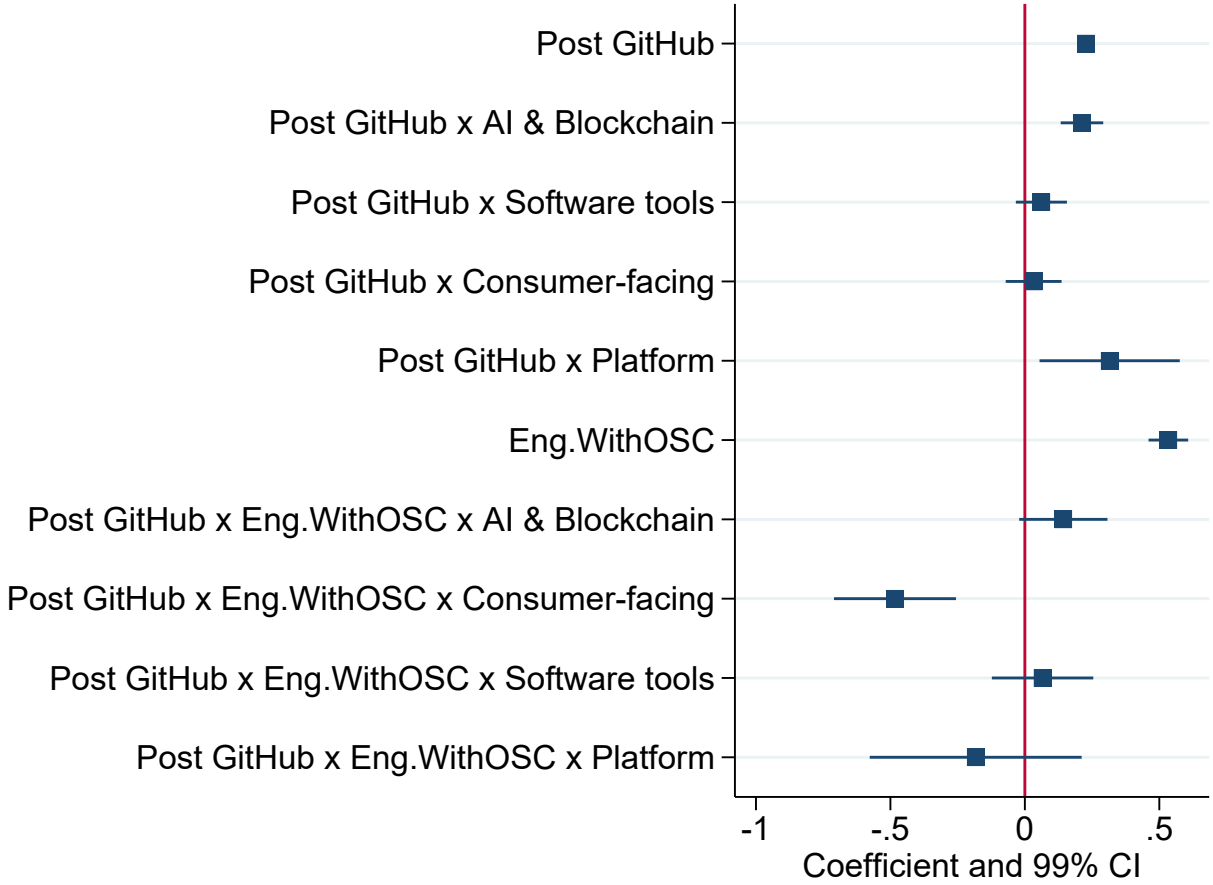
Notes: This figure shows how the probability that a startup will have raised a round changes with the number of engagements with OSCs by GitHub-active startups after GHE was enabled on AWS. The vertical lines represent 99% confidence intervals.

Figure 4: Correlations between startup domains and technology use-cases on GitHub



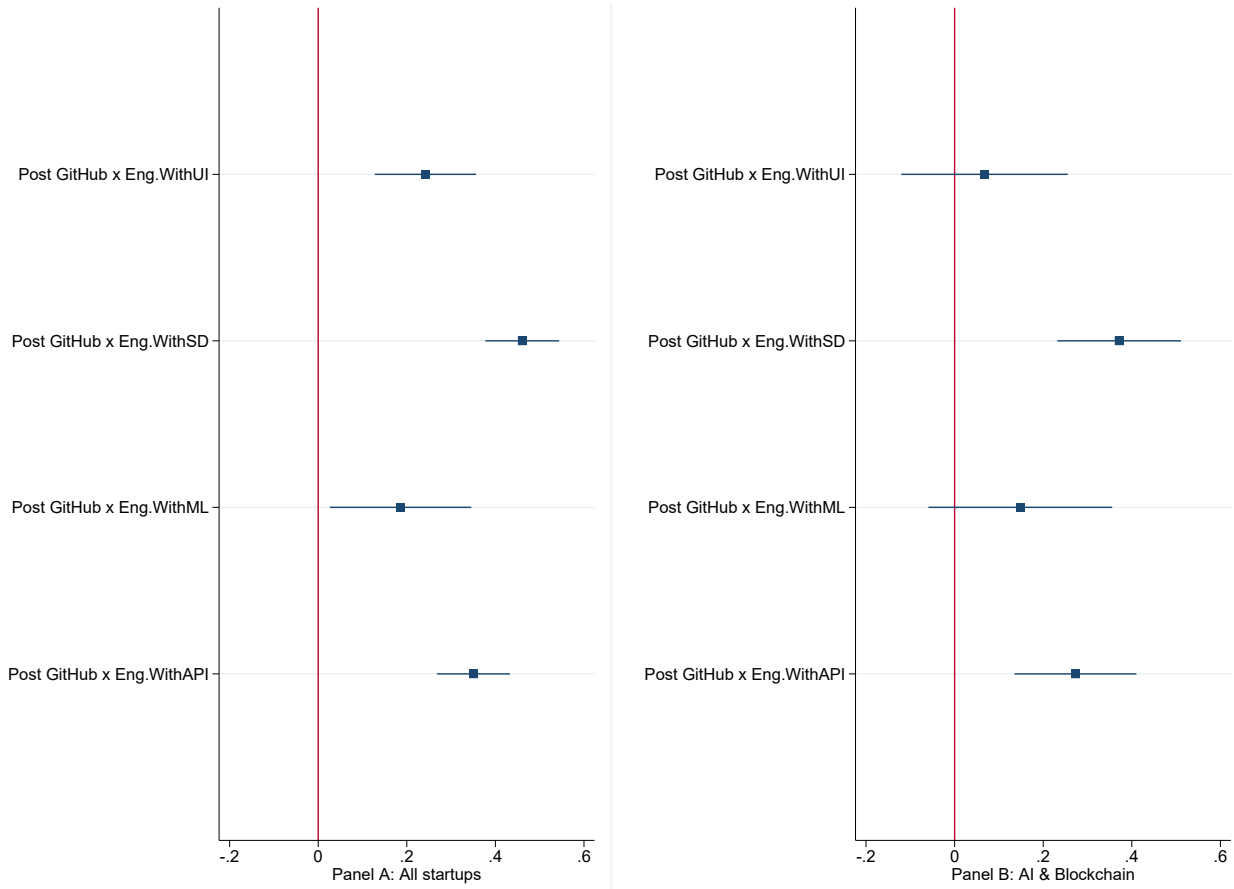
Notes: In this figure, we report the correlation coefficients and their 99% confidence intervals we obtain from estimating four cross-section models for the following outcomes: 1) a (0/1) indicator for whether a startup engaged with external repositories related to UI (Panel A); 2) a (0/1) indicator for whether a startup engaged with external repositories related to SD (Panel B); 3) a (0/1) indicator for whether a startup engaged with external repositories related to ML (Panel C); 4) a (0/1) indicator for whether a startup engaged with external repositories related to API (Panel D). The dependent variables of interest are: 1) an indicator identifying Platform startups; 2) an indicator identifying AI & Blockchain startups; 3) an indicator identifying Software Tool startups; and 4) an indicator identifying Consumer-facing startups. These are mutually exclusive categories; startups developing neither of these technologies represent the reference outcome. The reported coefficients are standardized by the means of the outcomes. In the regressions, we include treated-control group and startup founding year fixed effects. Standard errors are clustered by treated-control group.

Figure 5: Relationship between engaging with OSCs and funding: By startup domain



Notes: In this figure, we report the coefficients and respective 99% confidence intervals obtained from estimating a variant of Eq. (1) for the likelihood that a startup will have raised a first financing round by quarter t. Specifically, we interact $PostGitHub_{gt}$ and $EngagementWithOSC_i$ and $PostGitHub_{gt}$ with (0/1) indicators identifying: 1) platform startups; 2) AI & blockchain startups; 3) software tools startups; and 4) consumer-facing startups. In the regression, we include startup, year-quarter by domain, and treated-control group fixed effects. The startup domains are: Software Tools, AI & Blockchain, Platform, and Consumer-facing. The reported coefficients are standardized by the outcome mean. Standard errors are clustered by treated-control group.

Figure 6: Relationship between engaging with OSCs and funding: By technology use-cases



Notes: In this figure, we report the coefficients and respective 99% confidence intervals obtained from estimating Eq. (1) for the likelihood that a startup will have raised a first financing round by quarter t. We distinguish startups' engagements with external repositories related to: UI, SD, ML, API, and other. In Panel A, we consider all startups, while in Panel B, we zoom in on AI & Blockchain startups. In the regression, we include startup, year-quarter by startup domain, and treated-control group fixed effects. We only report the coefficients of interest, which are standardized by the outcome mean. Standard errors are clustered by treated-control group.

Table 1a: Summary statistics - Full sample of startups

	Mean	Std. Dev.	Min	Max	p50
Raised funds	0.357	0.479	0	1	0
Launched product	0.061	0.239	0	1	0
IPO	0.012	0.107	0	1	0
Acquired	0.081	0.273	0	1	0
GitHub	0.093	0.29	0	1	0
Engaged with external GitHub repos	0.084	0.28	0	1	0
Top Team	0.002	0.039	0	1	0
AI	0.039	0.194	0	1	0
Data Analytics	0.098	0.297	0	1	0
Information Technology	0.182	0.386	0	1	0
Internet Services	0.200	0.400	0	1	0
Software	0.356	0.479	0	1	0
N. Industry Groups	3	2	1	19	3
Software share	0.464	0.387	0	1	.5
California	0.296	0.457	0	1	0
Massachusetts	0.045	0.207	0	1	0
New York	0.128	0.334	0	1	0

Notes: In the case of product launches, we restrict the sample to companies founded after 2011, given that Product Hunt started only in 2014.

Table 1b: Summary statistics - Startups that engaged with an external repository on GitHub

	Obs	Mean	Std. Dev.	Min	Max	p50
SD/BE	13413	0.646	0.478	0	1	0
ML	13413	0.135	0.341	0	1	0
API	13413	0.628	0.483	0	1	0
UI	13413	0.277	0.448	0	1	0

Notes: The classification of repositories external to an organization's account was derived from implementing the machine learning algorithm described in the main text and in the Appendix.

Table 2a: Summary statistics - Matched sample

	Mean	Std. Dev.	Min	Max	p50
Raised funds	0.384	0.486	0	1	0
Launched product	0.087	0.282	0	1	0
IPO	0.006	0.077	0	1	0
Acquired	0.077	0.267	0	1	0
Engaged with external GitHub repos	0.21	0.407	0	1	0
Top Team	0.119	0.324	0	1	0
AI	0.071	0.257	0	1	0
Data Analytics	0.16	0.367	0	1	0
Information Technology	0.273	0.446	0	1	0
Internet Services	0.263	0.44	0	1	0
Software	0.545	0.498	0	1	1
N. Industry Groups	3	2	1	14	3
Software share	0.667	0.34	0	1	0.714
California	0.396	0.489	0	1	0
Massachusetts	0.047	0.212	0	1	0
New York	0.155	0.362	0	1	0

Notes: In the case of product launches, we restrict the sample to companies founded after 2011, given that Product Hunt started only in 2014.

Table 2b: Descriptive statistics by treated and control startups

	(1) Treatment=1		(2) Treatment=0		(3)
	Mean	S.D.	Mean	S.D.	Diff. (p-value)
Raised round	0.4976	0.5000	0.3540	0.4782	0.1436 (0.00)
Raised first round from VC	0.2356	0.4244	0.0994	0.2992	0.1362 (0.00)
Raised first round from successful investor	0.1797	0.3840	0.0683	0.2522	0.1114 (0.00)
Raised large financing amount	0.1392	0.3461	0.0633	0.2436	0.0758 (0.00)
Launched a product	0.1370	0.0046	0.0492	0.0015	0.0878 (0.00)
Observations	7207		27091		34298

Notes: This table reports descriptive statistics distinguishing between startups that had engaged with external repositories on GitHub (our measure for engagement with OSCs) and startups with no such engagement. Successful investors are those with a number of portfolio exits larger than the median. Portfolio exits are IPOs and acquisitions by portfolio startups in the five years prior to an investor's investment in startup i. A large financing amount is an amount greater than the median. The *Launched a product* variable is defined for startups founded after 2011.

Table 3: Raising a financing round

	Raising a first round			
	(1)	(2)	(3)	(4)
Post GitHub$_{gt}$	0.0518			
	(0.00244)			
Post GitHub$_{gt}$ × EngagementWithOSC$_i$	0.147	0.129	0.120	0.0811
	(0.00521)	(0.00534)	(0.00529)	(0.00460)
Startup FE	Y	Y	Y	Y
Yr-Quarter FE	Y			
Treated-control Group FE	Y			
Yr-Quarter × Treated-control Group FE		Y	Y	Y
Yr-Quarter × Industry Group FE			Y	Y
Startup Age × Startup FE				Y
Observations	778594	773769	773755	751674
R2	0.673	0.747	0.753	0.954
Mean D.V.	0.226	0.226	0.226	0.224

Notes: This table reports the results from estimating the difference-in-differences model described by Eq. (1) for the likelihood that a startup will have raised a first financing round by quarter t. We censor the sample on 2020q2. We observe each startup for up to 24 quarters, depending on whether the upper sample limit of 2020q2 is binding or not. $PostGitHub_{gt}$ is a time varying binary indicator that becomes one for all the startups in a treated-control group g after startup i engages in an external activity on GitHub. $EngagementWithOSC_i$ is our treatment indicator, which takes a value of one if startup i engages in at least one public activity across external repositories during the period we observe. It is our measure for engagement with OSCs on GitHub. In column 1, we report the results having included startup and year-quarter fixed effects, and fixed effects for each treated-control group g. In column 2, we add treated-control group by year-quarter fixed effects. In column 3, we additionally include industry group by year-quarter fixed effects. In column 4, we add startup by age (measured in years) fixed effects. Standard errors (in parentheses) are clustered by treated-control group.

Table 4: Engagement with external versus internal repositories

	(1)
	Raising a first round
Post GitHub$_{gt}$ × EngagementWithOSC$_i$	0.157
	(0.00535)
Post GitHub$_{gt}$ × Internal Activity$_i$	0.067
	(0.0051)
Startup FE	Y
Yr-Quarter × Treated-control Group FE	Y
Yr-Quarter × Industry Group FE	Y
Observations	946389
R2	0.759
Mean D.V.	0.261

Notes: This table reports the results from estimating a variant of the difference-in-differences model described by Eq. (1) for the likelihood that a startup will have raised a first financing round by quarter t. Here, we consider two treatments: (1) whether a startup engaged with an external repository ($EngagementWithOSC_i$), and (2) whether the startup engaged with an internal repository, which it directly controls ($InternalActivity_i$). $PostGitHub_{gt}$ is now a time varying binary indicator that becomes one for all the startups in a treated-control group g after startup i engages in an (internal or external) activity on GitHub. Control startups are those with no GitHub activity. Standard errors (in parentheses) are clustered by treated-control group.

Table 5: Raising a financing round - By technology novelty and the level of market competition

	Raising a first round	
	(1)	(2)
Post GitHub$_{gt}$ × EngagementWithOSC$_i$	0.0717	0.138
	(0.00774)	(0.00563)
Novel$_i$× Post GitHub$_{gt}$	0.0140	
	(0.00527)	
Novel$_i$× Post GitHub$_{gt}$ × EngagementWithOSC$_i$	0.0777	
	(0.0109)	
High Competition$_{it}$ × EngagementWithOSC$_i$		0.0558
		(0.0321)
High Competition$_{it}$ × Post GitHub$_{gt}$		-0.0358
		(0.00658)
High Competition$_{it}$ × Post GitHub$_{gt}$ × EngagementWithOSC$_i$		-0.143
		(0.0157)
Startup FE	Y	Y
Yr-Quarter × Treated-control Group FE	Y	Y
Yr-Quarter × Novel	Y	
Yr-Quarter × High Competition		Y
Observations	773769	773767
R2	0.751	0.749
Mean D.V.	0.226	0.226

Notes: This table reports the results from estimating a variant of the difference-in-differences model described by Eq. (1) for the likelihood that a startup will have raised a first financing round by quarter t. In column 1, we assess whether relationships uncovered in Table 3 vary depending on the novelty of a startup's technology. We measure technology novelty by whether the combination of a startup's industry group keywords at the startup's founding year is relatively new. That is, less than three years should have passed since a keyword combination first appeared on Crunchbase. In column 2, we assess whether the relationships uncovered in Table 3 vary by the level of market competition. We measure market competition by the number of startups active in a given year-quarter and with the same industry group keyword combination as the focal one. The market in which a focal startup operates is thus considered competitive if the number of startups possessing the same industry group keyword combination is greater than the 90^{th} percentile.

Table 6: Raising a financing round - Exploiting the integration of GHE with AWS

	Raising a first round	
	(1)	(2)
EngagementWithOSC$_{it}$	0.000124	0.000105
	(0.0000541)	(0.0000612)
Post AWS$_t$ × EngagementWithOSC$_{it}$	0.00363	0.00217
	(0.00141)	(0.00116)
Startup FE	Y	Y
Yr-Quarter FE	Y	Y
Observations	107636	75486
R2	0.718	0.766
Mean D.V.	0.289	0.291
Startup FE	Y	Y
Yr-Quarter × Industry Group FE	Y	Y
Observations	107636	75486
R2	0.710	0.761
Mean D.V.	0.289	0.291

We exploit the integration of GHE with AWS as a source of exogenous variation in the relative costs of engaging with the open source. In practice, estimate a modified version of Eq. (1) for the subsample of treated startups, interacting a startup's number of engagements with OSCs on GitHub in t with an indicator identifying the period after 2014q4 – when GHE was enabled on AWS ($PostAWS_t$). We include company and quarter fixed effects to control for fixed characteristics across startups and time effects. In column 1, we restrict the sample period to 2012q1-2017q4. In column 2, we restrict the sample period to 2013q1-2016q4. We impose these sample restrictions to focus on the period just before and after the integration of GHE with AWS. Standard errors (in parentheses) are clustered by startup.

Table 7: Raising a financing round - Exploiting the integration of GHE with AWS and distinguishing by technology novelty and the level of market competition

	Raising a first round	
	(1)	(2)
$\text{EngagementWithOSC}_{it}$	0.000116	0.000133
	(0.0000347)	(0.0000629)
Post $\text{AWS}_t \times \text{EngagementWithOSC}_{it}$	0.000627	0.00388
	(0.000963)	(0.00156)
$\text{EngagementWithOSC}_{it} \times \text{Novel}_i$	0.0000105	
	(0.0000692)	
Post $\text{GitHub}_{gt} \times \text{EngagementWithOSC}_{it} \times \text{Novel}_i$	0.00686	
	(0.00241)	
$\text{EngagementWithOSC}_{it} \times \text{High Competition}_{it}$		-0.000120
		(0.0000640)
Post $\text{GitHub}_{gt} \times \text{EngagementWithOSC}_{it} \times \text{High Competition}_{it}$		-0.00402
		(0.00196)
Startup FE	Y	Y
Yr-Quarter × Novel	Y	
Yr-Quarter × High Competition		Y
Observations	107636	107636
R2	0.718	0.715
Mean D.V.	0.289	0.289

We exploit the integration of GHE with AWS as a source of exogenous variation in the relative costs of engaging with OSCs on GitHub. We modify the model estimated in Table 5 to assess whether the relationship between a startup's engagement with OSCs and the likelihood of raising funds after GHE was enabled on AWS varies depending on the novelty of a startup's technology novelty and the level of market competition. We restrict the sample period to 2012q1-2017q4. Standard errors (in parentheses) are clustered by startup.

Table 8: Technology development or *just* signaling?

	(1)
	Raising a first round
Post GitHub$_{gt}$ × EngagementWithOSC$_i$	0.157
	(0.00535)
Post GitHub$_{gt}$ × Created/Modified Readme$_i$	0.008
	(0.0271)
Post GitHub$_{gt}$ × Other Internal Activity$_i$	0.068
	(0.0051)
Startup FE	Y
Yr-Quarter × Treated-control Group FE	Y
Yr-Quarter × Industry Group FE	Y
Observations	946389
R2	0.800
Mean D.V.	0.261

Notes: This table reports the results from estimating a similar model as the one reported in Table 5, decomposing internal activities on GitHub into creation or modification of readme files and the reminder. As in Table 5, $PostGitHub_{gt}$ is a time varying binary indicator that becomes one for all the startups in a treated-control group g after startup i engages in an (internal or external) activity on GitHub. Control startups are those with no GitHub activity. Standard errors (in parentheses) are clustered by treated-control group.

Table 9: Launching a product

	Launching a product			
	(1)	(2)	(3)	(4)
Post GitHub$_{gt}$	-0.0102			
	(0.00142)			
Post GitHub$_{gt}$ × EngagementWithOSC$_i$	0.0800	0.0766	0.0753	0.0218
	(0.00423)	(0.00424)	(0.00425)	(0.00260)
Startup FE	Y	Y	Y	Y
Yr-Quarter FE	Y			
Treated-control Group FE	Y			
Yr-Quarter × Treated-control Group FE		Y	Y	Y
Yr-Quarter × Industry Group FE			Y	Y
Startup Age × Startup FE				Y
Observations	574519	571572	571572	556579
R2	0.631	0.700	0.703	0.958
Mean D.V.	0.044	0.044	0.044	0.043

Notes: This table reports the results from estimating the difference-in-differences model described by Eq. (1) for the likelihood that a startup will have launched a product on Product Hunt by quarter t. As Product Hunt only started in 2014, we restrict our sample to those startups founded after 2011 for this analysis, given that older startups might have already launched their products, but without relying on the Product Hunt platform. $PostGitHub_{gt}$ is a time varying binary indicator that becomes one for all the startups in a treated-control group g after startup i engages with OSCs on GitHub. $EngagementWithOSC_i$ is our treatment indicator, which takes a value of one if startup i engages in at least one public activity across external repositories during the period we observe. It is our measure for engagement with OSCs. In column 1, we report the results having included startup and year-quarter fixed effects, and fixed effects for each treated-control group g. In column 2, we add treated-control group by year-quarter fixed effects. In column 3, we additionally include industry group by year-quarter fixed effects. In column 4, we add startup by age (measured in years) fixed effects. Standard errors (in parentheses) are clustered by treated-control group.

Table 10: Heterogeneity in financing outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
	High amount	Low amount	VC	Non-VC investor	Successful investor	Less successful investor
Post GitHub$_{gt}$ × EngagementWithOSC$_i$	0.0721	0.0483	0.105	0.0155	0.0892	0.0312
	(0.00381)	(0.00495)	(0.00435)	(0.00463)	(0.00403)	(0.00485)
Startup FE	Y	Y	Y	Y	Y	Y
Yr-Quarter × Treated-control Group FE	Y	Y	Y	Y	Y	Y
Yr-Quarter × Industry Group FE	Y	Y	Y	Y	Y	Y
Observations	773755	773755	773755	773755	773755	773755
R2	0.701	0.754	0.732	0.745	0.739	0.743
Mean D.V.	0.0502	0.1756	0.0731	0.1527	0.0572	0.1685

Notes: This table reports the results from estimating the difference-in-differences model described by Eq. (1). In column 1, we examine the likelihood that a startup will have raised a large financing round by quarter t. In column 2, we examine the likelihood that a startup will have raised a smaller financing round by quarter t. A large funding amount equals to one if the amount raised falls in the last quartile for the amount raised, a *Low amount* equals one for all others. In column 3, we examine the likelihood that a startup will have raised a first VC-led round. In column 4, we examine the likelihood that a startup will have raised a first non-VC-led round. In column 5, we examine the likelihood that a startup will have raised a first round from successful investors. In column 6, we examine the likelihood that a startup will have raised a first round from less successful investors. *Successful investors* are those with a number of exits in the five years prior to investing in startup i that falls in the last quartile of the distribution. Standard errors (in parentheses) are clustered by treated-control group.

Online Appendix for:

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

Table A1: Raising a financing round

	Raising a first round			
	(1)	(2)	(3)	(4)
Post GitHub$_{gt}$	0.0535			
	(0.00239)			
Post GitHub$_{gt}$ × EngagementWithOSC$_i$	0.149	0.131	0.121	0.0836
	(0.00516)	(0.00530)	(0.00524)	(0.00460)
Startup FE	Y	Y	Y	Y
Yr-Quarter FE	Y			
Treated-control Group FE	Y			
Yr-Quarter × Treated-control Group FE		Y	Y	Y
Yr-Quarter × Industry Group FE			Y	Y
Startup Age × Startup FE				Y
Observations	798900	794227	794224	771637
R2	0.673	0.747	0.753	0.954
Mean D.V.	0.229	0.229	0.228	0.226

Notes: This table reports the results from estimating the difference-in-differences model described by Eq. (1) for the likelihood that a startup will have raised a first financing round by quarter t. This time, we do not match treated and control startups based on *Top Team*. We censor the sample on 2020q2. We observe each startup for up to 24 quarters, depending on whether the upper sample limit of 2020q2 is binding or not. $PostGitHub_{gt}$ is a time varying binary indicator that turns to one for all the startups in a treated-control group g after startup i engages with OSCs on GitHub. $EngagementWithOSC_i$ is our treatment indicator, which takes a value of one if startup i engages in at least one public activity across external repositories during the period we observe. It is our measure for engagement with OSCs. In column 1, we report the results having included startup and year-quarter fixed effects, and fixed effects for each treated-control group g. In column 2, we add treated-control group by year-quarter fixed effects. In column 3, we additionally include industry group by year-quarter fixed effects. In column 4, we add startup by age (measured in years) fixed effects. Standard errors (in parentheses) are clustered by treated-control group.

Table A2: Raising a financing round

	Raising a first round			
	(1)	(2)	(3)	(4)
Post GitHub$_{gt}$	0.0245			
	(0.00323)			
Post GitHub$_{gt}$ × EngagementWithOSC$_i$	0.200	0.185	0.180	0.0558
	(0.00479)	(0.00498)	(0.00495)	(0.00436)
Startup FE	Y	Y	Y	Y
Yr-Quarter FE	Y			
Treated-control Group FE	Y			
Yr-Quarter × Treated-control Group FE		Y	Y	Y
Yr-Quarter × Industry Group FE			Y	Y
Startup Age × Startup FE				Y
Observations	640658	634145	634140	615462
R2	0.683	0.783	0.786	0.956
Mean D.V.	0.308	0.304	0.307	0.304

Notes: This table reports the results from estimating the difference-in-differences model described by Eq. (1) for the likelihood that a startup will have raised a first financing round by quarter t. This time, we impose that the startups in the control group (up to five) are founded in the same state and year as the treated startups and share at least one technology keyword with the treated startup. We censor the sample on 2020q2. We observe each startup for up to 24 quarters, depending on whether the upper sample limit of 2020q2 is binding or not. $PostGitHub_{gt}$ is a time varying binary indicator that becomes one for all the startups in a treated-control group g after startup i engages with OSCs on GitHub. $EngagementWithOSC_i$ is our treatment indicator, which takes a value of one if startup i engages in at least one public activity across external repositories during the period we observe. It is our measure for engagement with OSCs. In column 1, we report the results having included startup and year-quarter fixed effects, and fixed effects for each treated-control group g. In column 2, we add treated-control group by year-quarter fixed effects. In column 3, we additionally include industry group by year-quarter fixed effects. In column 4, we add startup by age (measured in years) fixed effects. Standard errors (in parentheses) are clustered by treated-control group.

Table A3: Launching a high-quality product

	Launching a high-quality product			
	(1)	(2)	(3)	(4)
Post GitHub$_{gt}$	0.000884			
	(0.000245)			
Post GitHub$_{gt}$ × EngagementWithOSC$_i$	0.00323	0.00316	0.00318	0.00218
	(0.000700)	(0.000705)	(0.000710)	(0.000782)
Startup FE	Y	Y	Y	Y
Yr-Quarter FE	Y			
Treated-control Group FE	Y			
Yr-Quarter × Treated-control Group FE		Y	Y	Y
Yr-Quarter × Industry Group FE			Y	Y
Startup Age × Startup FE				Y
Observations	574519	571572	571572	556579
R2	0.312	0.436	0.438	0.898
Mean D.V.	0.0011766	0.0011827	0.0011827	0.0011984

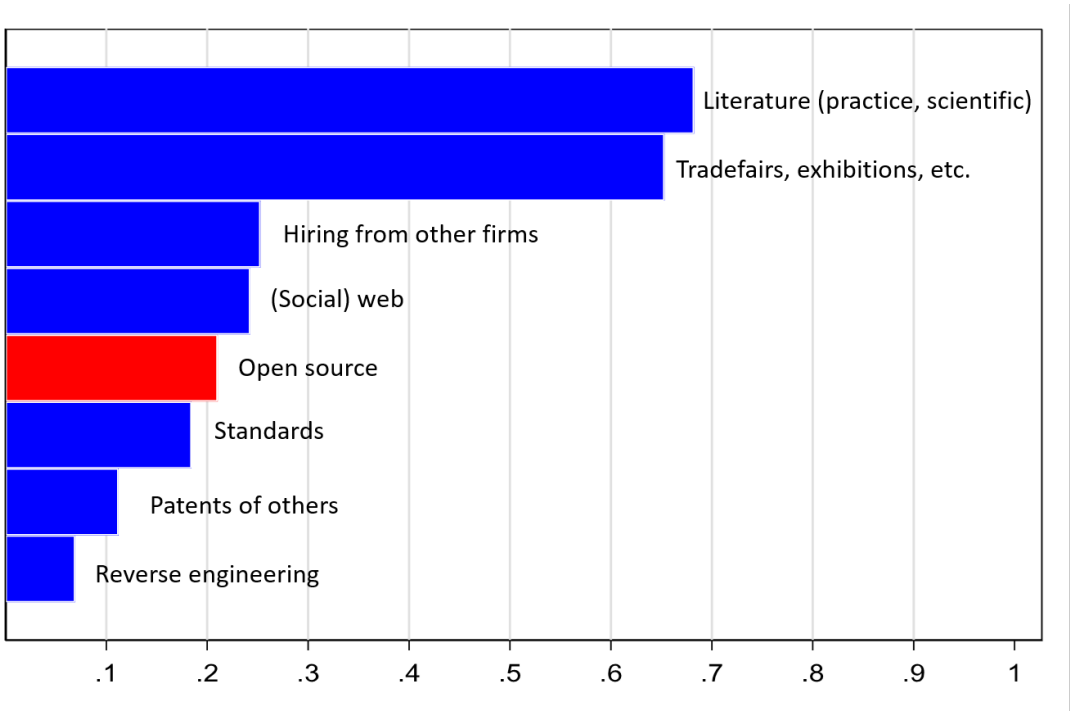
Notes: This table reports the results from estimating the difference-in-differences model described by Eq. (1) for the likelihood that a startup will have launched a high-quality product on Product Hunt by quarter t. A high-quality product is one whose number of ratings or upvotes received as of September 2023 is greater than the 75 percentile. As Product Hunt only started in 2014, we restrict our sample to those startups founded after 2011 for this analysis, given that older startups might have already launched their products, but without relying on the Product Hunt platform. $PostGitHub_{gt}$ is a time varying binary indicator that becomes one for all the startups in a treated-control group g after startup i engages with OSCs on GitHub. $EngagementWithOSC_i$ is our treatment indicator, which takes a value of one if startup i engages in at least one public activity across external repositories during the period we observe. It is our measure for engagement with OSCs. In column 1, we report the results having included startup and year-quarter fixed effects, and fixed effects for each treated-control group g. In column 2, we add treated-control group by year-quarter fixed effects. In column 3, we additionally include industry group by year-quarter fixed effects. In column 4, we add startup by age (measured in years) fixed effects. Standard errors (in parentheses) are clustered by treated-control group.

Table A4: Launching a first product

	Launching a first product			
	(1)	(2)	(3)	(4)
Post GitHub$_{gt}$	0.00101			
	(0.000238)			
Post GitHub$_{gt}$ × EngagementWithOSC$_i$	0.00271	0.00267	0.00269	0.00211
	(0.000667)	(0.000675)	(0.000675)	(0.000786)
Startup FE	Y	Y	Y	Y
Yr-Quarter FE	Y			
Treated-control Group FE	Y			
Yr-Quarter × Treated-control Group FE		Y	Y	Y
Yr-Quarter × Industry Group FE			Y	Y
Startup Age × Startup FE				Y
Observations	574519	571572	571572	556579
R2	0.321	0.444	0.446	0.903
Mean D.V.	0.0011523	0.0011582	0.0011582	0.0011786

Notes: This table reports the results from estimating the difference-in-differences model described by Eq. (1) for the likelihood that a startup will have launched the first version of a product on Product Hunt by quarter t. As Product Hunt only started in 2014, we restrict our sample to those startups founded after 2011 for this analysis, given that older startups might have already launched their products, but without relying on the Product Hunt platform. $PostGitHub_{gt}$ is a time varying binary indicator that becomes one for all the startups in a treated-control group g after startup i engages with OSCs on GitHub. $EngagementWithOSC_i$ is our treatment indicator, which takes a value of one if startup i engages in at least one public activity across external repositories during the period we observe. It is our measure for engagement with OSCs. In column 1, we report the results having included startup and year-quarter fixed effects, and fixed effects for each treated-control group g. In column 2, we add treated-control group by year-quarter fixed effects. In column 3, we additionally include industry group by year-quarter fixed effects. In column 4, we add startup by age (measured in years) fixed effects. Standard errors (in parentheses) are clustered by treated-control group.

Figure A1: Firms reliance on external sources of knowledge



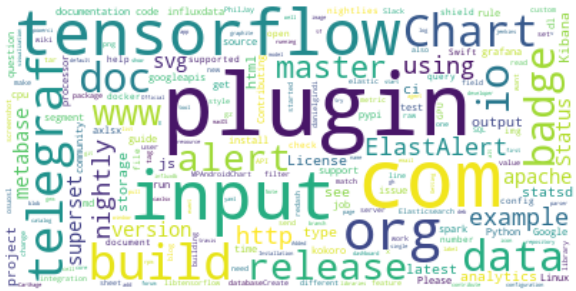
Notes: The data used to construct this graph comes from the Mannheim Innovation Panel, a representative survey of 4909 German firms conducted by the Center for European Economic Research (https://www.zew.de/PJ345). The numbers depicted relate to the question on what sources firms used to access knowledge from external sources in the years 2016-2018.

Figure A2: Repository classification: word clouds

Software (Back end), I

Software (Back end), II

Machine Learning



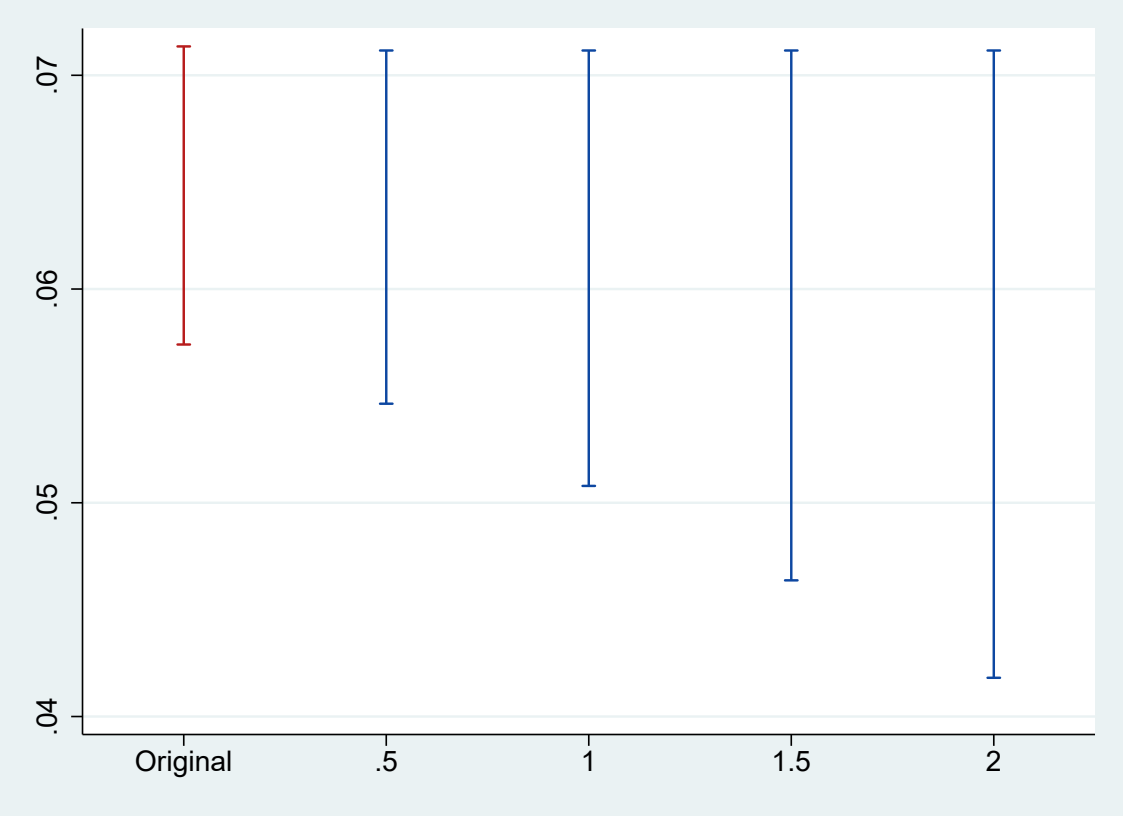
API

User Interface (UI)

Human Resources (HR)

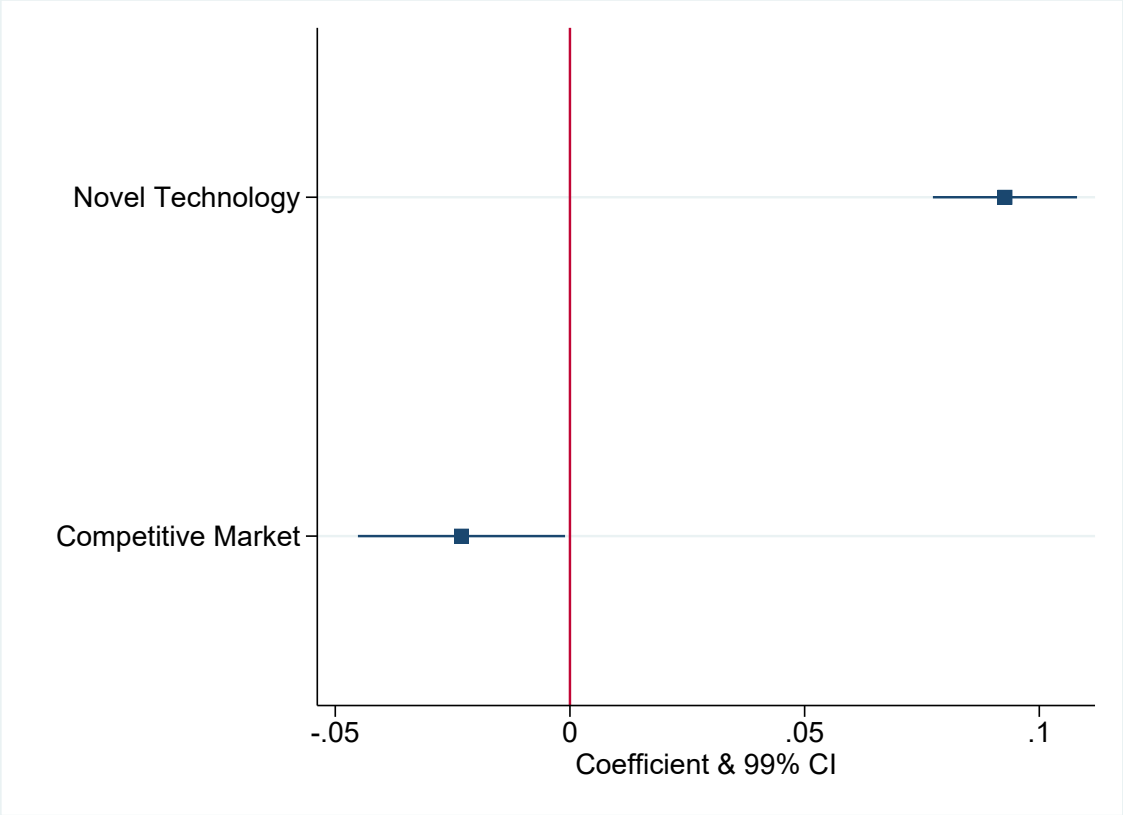
Notes: This figure displays the most common words for each of the technology use-cases we have considered. Note that these word clouds are based on the full classified sample, for which we used the hand-coded training sample as starting point for k-nearest neighbors. To construct this figure, word counts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

Figure A3: Sensitivity analysis



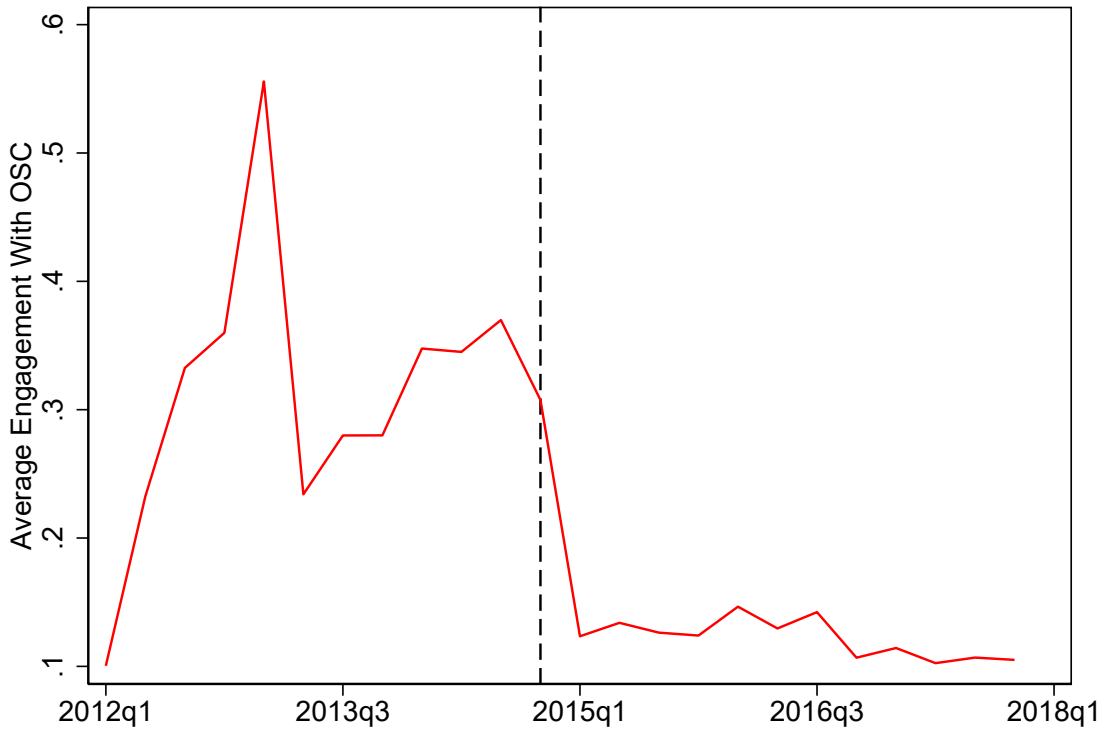
Notes: This figure plots (1) the confidence interval for the coefficient of the first quarter after treatment obtained from the specification in column 1 of Table 3 (red bar) and (2) fixed length confidence intervals (blue bars) when Δ, that is, the set of possible differences in trends, is $\Delta = \Delta^{SD}(M)$ for different values of M. The latter represents the amount by which the slope of δ can change between consecutive periods.

Figure A4: Becoming engaged with OSCs on GitHub: By startup technology novelty and the level of market competition



Notes: In this figure, we report the coefficients and their 99% confidence intervals obtained from estimating a cross-section model for the likelihood that a startup becomes engaged with OSCs on GitHub as a function of whether a startup's technology is novel and the level of market competition. We measure technology novelty by whether the combination of a startup's industry group keywords at the startup's founding year is relatively new. That is, less than three years should have passed since a keyword combination first appeared on Crunchbase. To operationalize the level of competition in each technology space, we count the number of startups active in a given year-quarter and with the same industry group keywords as the focal one. The market in which a focal startup operates is thus considered competitive if the number of startups possessing the same industry group keyword combination is greater than the 90th percentile. We include startup founding year and treated-control group fixed effects. Standard errors are clustered by treated-control group.

Figure A5: The deployment of GHE on AWS



Notes: This figure displays the average startup engagement with OSCs on GitHub. The vertical dotted line represents the quarter in which GHE was deployed on AWS.

A Repository classification

We classify the public repositories of all organizations, as well as the external repositories with which the organizations interact through commits, pull requests, or forks according to their type. We distinguish between repositories that pertain to Software Development/Backend, Machine Learning, Application Programming Interface, and User Interface. 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 a 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 by 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.[24]

[24] 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.