Corporate Adoption of an Open Innovation Strategy: Evidence from GitHub

Braiden Coleman	Karson Fronk	Kristen Valentine
University of Georgia	Texas Christian University	University of Georgia
braiden.coleman@uga.edu	*karson.fronk@tcu.edu*	*kristen.valentine@uga.edu*

June 2025

ABSTRACT

An open innovation strategy involves sharing knowledge and soliciting external feedback to solve problems and create new technologies. We examine the choice to adopt an open innovation strategy and the potential benefits for adopting firms, using firms' online GitHub postings of software projects as a setting. We find that firms choose open innovation for a portion of their innovation portfolio as a complement to patenting and trade secrecy. Results also show a significantly positive association between GitHub activity and future earnings that is increasing in the amount of crowdsourced feedback and the extent to which peers use the posted technology. Further, GitHub adopting firms experience greater returns to R&D investment, suggestive of a managerial learning channel. However, GitHub usage is also positively associated with future cybersecurity breaches. Using project-level announcement returns to corroborate our firm-year earnings results, we find that firms experience positive, risk-adjusted market returns upon posting projects to GitHub. This pattern of results is consistent with an open innovation strategy being net beneficial for adopting firms even as it presents some costs. Overall, the evidence advances our understanding of which firms choose an open innovation strategy and the potential sources of private value creation from adoption, which has implications for the role of knowledge sharing in the new economy.

Keywords: open innovation, GitHub, crowdsourcing, managerial learning

JEL codes: M40, M41, O31, O32, O33, O36

We thank Abigail Allen, Thomas Bourveau, John Campbell, Ted Christensen, Travis Dyer, Nathan Goldman, Hengda Jin (discussant), Jinhwan Kim, Tim Martens, Jeff McMullin, Ken Merkley, Brian Miller, Peter Pope, Karl Schuhmacher, Bridget Stomberg, and Ane Tamayo (discussant) for insightful comments. We thank workshop participants at Harvard University and the University of Georgia, the BYU Accounting Research Symposium, the London Business School Accounting Conference, and the Texas A&M Accounting Research Conference for valuable feedback and suggestions. We are also grateful to representatives at GitHub for helpful discussions.

"The software industry continues to grow at a massive clip. More and more traditional companies are realizing that to compete and grow in a digital world, they must look, think, and act like software companies themselves." -McKinsey & Company, 2022

1 INTRODUCTION

Across a portfolio of intellectual property, firms often choose to protect their innovations through trade secrecy or patenting depending on the underlying technology. However, we document an emerging trend in which firms share knowledge with external parties to innovate and solve problems, a strategy known as open innovation (Chesbrough 2004). Specifically, we examine firms' disclosure of software source code on the collaborative software development platform GitHub as a type of open innovation. While an open innovation strategy may subject a firm to proprietary costs and other threats (e.g., cybersecurity concerns), it can also benefit the adopting firm by facilitating direct improvements to the software, encouraging widespread adoption of the technology, and improving future project decisions. In this study, we investigate the extent to which firms engage in open innovation and seek to better understand how this strategy can potentially be beneficial.

A defining feature of innovation that drives long-term economic growth is non-excludability—the extent to which knowledge and ideas, once created, can be freely accessed and used by others without permission or payment (Romer, 1990). Unlike patents or trade secrets, which enforce at least partial excludability by limiting access to proprietary technologies, open innovation can promote socially beneficial knowledge transfers by reducing the costs for others to build on the adopting firms' technological advancements. In our setting, we suggest that firms' use of GitHub represents a form of open innovation wherein code and software projects are publicly shared and available for external use, modification, and redistribution. Understanding this phenomenon is particularly important in light of the growing significance of software to the modern firm. Indeed, it is estimated that over $1.2 trillion will be invested in software in 2025 alone (Gartner, 2024). Our study examines the choice to adopt an open innovation strategy as well as potential firm-level benefits and costs using the economically significant setting of collaborative software development.

While open innovation can facilitate knowledge “spill-outs”—enabling others to build upon a firm’s original work—it can also generate important knowledge “spill-ins,” as external contributors improve and extend the firm’s projects. The potential benefits to the adopting firm from knowledge spill-ins are one important focus of our study. We posit that open innovation via GitHub can benefit adopting firms in at least three ways. First, GitHub enables its users to receive suggestions on publicly posted software projects (also termed repositories) from a vast, worldwide network of external developers. External parties can submit coding suggestions for project administrators to review and determine whether to incorporate in the source code. Second, outside developers can also copy posted code for use in follow-on projects, facilitating widespread adoption of the posting firm’s technology. Third, the GitHub platform allows firm managers to observe community-level interest in the firm’s software projects (e.g., the number of code duplications), which enables managerial learning from the crowd. However, open innovation also has potential costs that could partially offset the expected benefits. Specifically, publicly posting source code could open the firm to proprietary costs or increased cybersecurity threats.

Because pursuing open innovation via GitHub is a strategic choice, we expect that managers, on average, accurately weigh the costs and benefits ex ante and pursue an open innovation strategy if it is net beneficial to the firm. Importantly, because open software development is a new technological advancement, some firms will be better positioned than others to benefit from this strategy (Romer, 1990). Thus, even if both GitHub adopting and non-adopting firms aim to maximize profits, we expect that firms who choose an open innovation strategy have capabilities that enable them to uniquely benefit from GitHub adoption in a way unavailable to non-adopting firms given their current capabilities.

To examine firms' open innovation activity, we create a unique dataset that links GitHub software projects to publicly traded firms. Given the novelty of our dataset, our empirical strategy first provides descriptive evidence on public firms' GitHub usage and explores a determinants model. After controlling for these determinants, we examine how GitHub usage associates with future earnings as a proxy for private value creation. We then investigate the channels through which open innovation can generate value and the costs that could reduce the net benefits of an open innovation strategy.

Reflective of the extensive use of software in the economy, we find that a broad cross-section of approximately 1,597 unique firms across all major industries adopt GitHub from 2008 to 2021. On average, firms that have public repositories on the platform release approximately 15 new projects each year. In addition, disclosed software projects receive 16 external, accepted contributions to the code, and the projects are copied 27 times, on average. We provide a detailed description of the data in Appendix A.

Turning to the determinants of GitHub adoption, we find that trade secrecy usage and the number of software patent filings are positively associated with GitHub adoption, suggesting that firms use open innovation as a complement to traditional forms of intellectual property protection across a portfolio of innovations. In terms of performance and the financial reporting environment, GitHub usage is positively associated with profitability, R&D and advertising expenditures, analyst following, size, and industry concentration.

We next examine the association between GitHub activity and value creation. We follow an approach implemented by Curtis, McVay, and Toynbee (2020), who explore the changing implications of research and development expenditures for future profitability. Following Curtis et al. (2020), we use aggregate future earnings as our primary dependent variable, with the expectation that the benefits of open innovation ultimately manifest as an increase in future revenue or a reduction in costs for adopting firms relative to control firms that do not similarly benefit.[1] Specifically, we regress firms' five-year aggregate future earnings on the number of new projects posted to GitHub in a fiscal year. We find a significantly positive association between new GitHub activity and future realized net income, which is consistent with open innovation creating firm value.[2] In terms of economic significance, we find that a one standard deviation increase in new GitHub activity is associated with an approximately 6.85% increase in earnings relative to the unconditional mean in our sample period.

We posit that GitHub usage can benefit the adopting firm through three non-mutually exclusive means: 1) directly improving the underlying software project through crowdsourced software development, 2) facilitating widespread adoption of the firm's technologies, and 3) providing a source of managerial learning that improves future project selection decisions. First, open innovation can accelerate the development of new products and aid with technology advancement (Chesbrough, 2004). Crowdsourced software development can directly improve the disclosed software and any related products—leading to an increase in revenues. External contributions could also potentially decrease software development costs, reducing overall expenses. After separating GitHub repositories into high- and low-crowdsourcing subsets, we find that firms whose projects have more crowdsourced suggestions have a stronger association with future earnings than those with fewer crowdsourced suggestions. This suggests that crowdsourcing software development is one channel through which open innovation can create value for firms.[3]

Second, pursuing open innovation can lead to more widespread adoption of the firms' technologies as others build on the disclosed work (Bonaccorsi and Rossi, 2003). Increased GitHub

[1] Ideally, we would observe project-level financial outcomes rather than firm-level, aggregated earnings. However, project-level revenue and expense data are unavailable, and there is no mapping between GitHub projects and either licensing or patent data. Thus, we rely on future earnings as the final realization of any benefit.

[2] As discussed in section 5.2, our results are robust to alternative approaches in measuring GitHub usage, separately using individual years of future earnings (e.g., t+1) instead of aggregated future earnings as a dependent variable, and implementing an alternative earnings model (Ciftci and Cready, 2011).

[3] For example, after Twitter posted a portion of its source code on GitHub, a developer from Argentina found a security vulnerability and alerted the company—improving Twitter's technology as a direct result of making the code public. The GitHub activity of many developers is viewed as a core component of their resume, which incentivizes quality contributions. See Appendix B for this and other examples of how firms can benefit from GitHub usage.

usage can benefit the adopting firm by expanding the number of products and services that use the software or by facilitating the sale of consulting and installation services. We proxy for widely adopted projects by exploiting variation in the number of copies project users make. We find that firms whose projects have more copies have a stronger association with future earnings, consistent with open innovation facilitating the dissemination and adoption of firms' technologies.

Third, managers can enhance their learning and improve firm-level investment decisions by soliciting market feedback to voluntary disclosures (Jayaraman and Wu, 2020). Similarly, the feedback on firms' public GitHub repositories can improve managers' information set, enabling managerial learning (Roychowdhury et al., 2019; Guo and Zhong, 2023). We regress firms' future five-year aggregate earnings on the interaction between R&D expenditures and GitHub activity. If GitHub firms experience greater returns to R&D investment, we anticipate the interaction to be positively associated with future earnings. The results confirm our expectations, suggesting that GitHub adoption contributes to improved project selection and/or resource allocation decisions.

We embrace the voluntary nature of choosing to adopt an open innovation strategy as a feature of our study. Indeed, one important objective of our study is to understand the firm-level determinants of pursuing an open innovation strategy. At the same time, this endogenous choice leaves open the possibility that the observed positive association between GitHub activity and future earnings is due to fundamental differences in adopting versus non-adopting firms rather than an open innovation strategy per se. We take several steps to triangulate inferences, though we caution the reader than our evidence should be interpreted as an important step toward understanding an important phenomenon rather than allowing for causal inference.

First, our models include the time-varying firm characteristics identified in our determinants model as control variables as well as industry and year fixed effects. Second, we employ a modified control function approach that is intended to control for endogeneity (Klein and Vella, 2010; Armstrong et al., 2022). More specifically, this approach addresses endogeneity by modeling and controlling for unobserved factors, including selection effects, that jointly influence both a firm's GitHub activity and its future earnings (Armstrong et al., 2022). Using this method, we continue to find a robust positive association between GitHub usage and future earnings, even after the inclusion of firm and year fixed effects. Third, we show that our findings are robust to the use of alternative control samples. For instance, we limit our sample to include only GitHub adopting firms, thereby holding constant a firm's choice to use GitHub and exploiting variation in the intensity of GitHub usage, and the results remain consistent. We also highlight that our tests examining the channels of value creation strengthen the mosaic of our evidence (Glaeser and Guay, 2017). Nonetheless, while the pattern of results supports the interpretation that GitHub usage creates value for adopting firms and we take steps to address selection effects, we acknowledge that our design inherently exploits conditional correlations rather than providing causal inference.

Although the result that GitHub usage is positively associated with future earnings suggests that open innovation is net beneficial to adopting firms, we next explore potential costs of this innovation strategy that may partially offset the benefits. Anecdotal evidence suggests that GitHub adoption could expose firms to cybersecurity risks.[4] To investigate this possibility, we obtain data on cybersecurity breaches for firms in our sample. We find a significantly positive association between firms' GitHub activity and the likelihood of future data breaches. This analysis is consistent with the notion that open innovation can generate certain downside costs for firms even if the overall impact is net positive.

To provide more granular evidence at the project level, we next examine investors' short-term market response to firms' GitHub software postings. While certain investors may be unaware of firms' GitHub activity, a review of sell-side analyst research reports suggests that some analysts

[4] For example, GitHub use can lead to the unintentional release of sensitive information, including API keys and passwords (see https://cybersecuritynews.com/39m-secret-api-keys-credentials-leaked-from-github/). Public code files can also reveal details about a firm's internal systems, such as infrastructure configurations and sensitive software dependencies (see https://www.techtarget.com/searchitoperations/news/366621078/GitHub-Actions-supply-chain-attack-spotlights-CI-CD-risks).

monitor firms' GitHub profiles (e.g., Bora, 2022). We find significantly positive risk-adjusted returns following GitHub project postings, consistent with investors viewing open innovation activities favorably. Further, the market reaction concentrates around more widely adopted projects. These findings complement the evidence from our earnings tests, suggesting that investors anticipate greater value for firms utilizing GitHub.[5]

We contribute to the literature on innovation in accounting by introducing open innovation as a complement to firms' other innovative pursuits. Prior work typically examines either trade secrecy usage or patenting as methods to preserve intellectual property rights (Glaeser, 2018; Glaeser and Lang, 2024). We examine one of the purest forms of non-excludable innovation—open-source software—and find that firms choose an open innovation strategy as a complement to patenting and trade secrecy usage. By doing so, we shed light on firms' strategic choice to invest in non-excludable innovation. Moreover, our findings suggest that firms appear to benefit from sharing information in the software setting, rather than suffering competitive harm. Although we caution that the benefits we observe for our sample firms would not necessarily generalize to firms subject to a hypothetical open innovation mandate, evidence supporting these channels of value creation for open innovating firms is new to the accounting literature.

Our work also relates to the literature on voluntary innovation disclosure.[6] While open innovation is a strategic operating choice, inherent to that choice is a disclosure commitment. We document that firms voluntarily reveal proprietary information via GitHub—with virtually no access restrictions and potentially at the cost of weakening intellectual property protection rights. In contrast to extant literature that examines the tradeoff between the proprietary costs of disclosure

[5] While these findings align with evidence from our earnings tests, we acknowledge that investors' positive short-term reactions could reflect a response to firm disclosures about in-process R&D revealed through GitHub postings. Nonetheless, the positive market response to GitHub postings is informative, as it suggests that investors i) monitor firms' GitHub activity and ii) do not, on average, view GitHub activity as value-destroying—given the observed positive reaction.

[6] Prior work examines innovation disclosures in the 10-K filing (Jones 2007; Merkley 2014), patent filings (Glaeser et al., 2020; Glaeser and Landsman 2021; Kim and Valentine 2021), scientific publications (Johnson 2014; Baruffaldi et al., 2024), contract redactions (Glaeser 2018), and product disclosures (Cao et al., 2018; Chu et al., 2025). Bourveau et al. (2022) examine GitHub postings as a form of other disclosures initial coin offering ventures make, but do not systematically evaluate GitHub as an open innovation platform employed by a broad sample of public firms.

and capital market benefits (Guo et al., 2004; Cao et al., 2018), our evidence suggests firms can experience real activity benefits from voluntarily releasing detailed innovation information. Given that the social value of an innovation depends on its public disclosure (Armstrong et al., 2024), understanding the sources of value creation from open innovation is key to stakeholders seeking to encourage the supply of innovation information. Further, while research finds that patent disclosures can inspire other firms' future innovations (i.e., knowledge spill-outs) (Dyer et al., 2024), our study highlights the potential benefits from external parties directly contributing to firms' ongoing innovations via the crowdsourcing aspect of GitHub (i.e., knowledge spill-ins) (Kim and Valentine, 2021).

Finally, our study contributes to the open innovation literature. We provide large-scale evidence on the determinants of open innovation for U.S. firms, which heretofore has been limited (West and Boger, 2017). This setting is significant as 90% of Fortune 100 companies develop software via the GitHub platform (GitHub, 2023). Furthermore, our study explores the *channels* through which open innovation influences firm performance, adding new insights. For example, as Piller and West (2014) suggest, much of the open innovation literature focuses on organizations as the key actors (e.g., strategic innovation alliances between firms or firms' purchases of intellectual property from other organizations). In our test of the channels of value creation, we examine the impact of individual developers' contributions to GitHub postings. Our setting thus facilitates knowledge flows between the firm as an *organization* and external *individuals* (i.e., GitHub contributors). As such, we provide evidence on firms' interactions with individual inventors (i.e., developers), which Glaeser and Lang (2024) identify as an emerging innovation research opportunity. To encourage future work, we make our dataset available to researchers.

2 BACKGROUND AND EMPIRICAL PREDICTION

2.1 Institutional Setting

Software is one of the fastest-growing investment segments in the world. GitHub, the largest software aggregation platform, launched in April 2008 and allows software developers "to build, scale, and deliver secure software" instantaneously. In 2023, approximately 100 million developers and four million organizations (including 90% of Fortune 100 companies) were active on the platform, which hosts a total of roughly 420 million repositories (GitHub, 2023). Individuals and organizations can post a variety of repositories on GitHub using different software packages and programming languages (e.g., Python, C++, Swift, Java, or PHP). The primary objective behind such posts is to 1) allow the community to contribute and improve the efficiency or execution of the owners' projects and 2) help other users advance their own projects by adopting the owners' posted software, either entirely or in part.

Coding suggestions on GitHub software projects can occur when the owner solicits help by reporting an "issue" or when contributors offer unsolicited suggestions (i.e., pull requests), which the repository owner must approve. Repository owners can also choose to post their repositories privately with access restricted to only those individuals within the organization or designated by the owner.[7] For public repositories, GitHub contributors can perform many actions that include submitting pull requests, bookmarking projects (i.e., stargazers), copying programs (i.e., forks), and offering comments. In anecdotal discussions, individual GitHub contributors view their GitHub profiles to be just as important as their work or education experience listed on a resume, if not more so, because the GitHub profile manifests their expertise. Thus, contributors have incentives to provide high-quality contributions on public repositories.

[7] Furthermore, firms can supplement their GitHub profile—or circumvent GitHub entirely—by providing their open-source projects on their own websites. However, we expect these alternative sources of open-source activity to bias against finding an effect in our setting.

The GitHub accounts we examine are those of the firm, not individual developer profiles. Firms post projects to GitHub under various licenses. While firms can choose from a variety of licenses, in untabulated analysis we find that more than 77 percent of the licenses employed in our sample are approved by the Open Source Initiative as open-source licenses that "allow software to be freely used, modified, and shared."[8] Of the remaining licenses, an inspection of license titles suggests that they generally do allow for freedom of use, even if they do not meet the Open Source Initiative's stricter criteria that requires a license review process. While the particulars of a given license may vary, at a minimum, publicly posting source code increases uncertainty about the enforceability of any related intellectual property rights. For example, publicly posting code could make successfully obtaining a patent more difficult as public GitHub postings constitute public disclosure and require firms to file for patent protection within twelve months (Laforgia, 2019). In short, publicly posting source code on GitHub increases uncertainty about the firm's ability to prevent competitors from using their technology independent of the type of license chosen. Thus, it constitutes "open innovation" relative to the baseline of not publicly posting source code.

2.2 Open Innovation

Open innovation "is a paradigm that assumes that firms can and should use external ideas as well as internal ideas, and internal and external paths to market, as firms look to advance their technology" (Chesbrough, 2006). Research on open innovation utilizes predominantly surveys and case studies. These studies hypothesize that open innovation provides firms with new knowledge, solutions to problems, and may accelerate product development, even as the type of information systems used and open innovation contests may provide challenges (Laursen and Salter, 2006; Scacchi et al., 2006; Majchrzak and Malhotra, 2013; Boudreau, 2010; Boudreau et al., 2011).

[8] For more information on the Open Source Initiative's list of compliant licenses and its definition of open source, see https://opensource.org/licenses/.

In terms of open innovation and firm performance, prior findings focus on specialized settings and document mixed results (Dahlander and Gann, 2010; Bigliardi et al., 2020; Lu and Chesbrough, 2022). For instance, using survey data on European firms, Greco et al. (2016) find that the number of categories of innovation sources that firms use is positively associated with the percentage of sales resulting from new products or services. However, evidence also suggests that a curvilinear association exists between open innovation activities and firm performance. Using data on manufacturing firms from the U.K. innovation survey, Laursen and Salter (2006) find a negative association between the squared number of external sources that firms rely on in their innovative activities and the fraction of a firm's sales relating to new products. Greco et al. (2016) document diminishing marginal performance effects of open innovation. Finally, other work suggests that open innovation can lead to negative outcomes. For instance, using survey data from Danish manufacturing firms, Knudsen and Mortensen (2011) find that project models with more 'openness' can lead to worse timing to market and slower, more costly product development.[9] In sum, using survey data and small sample evidence, evidence on the direction of the relation between open innovation and firm performance is mixed.

Our study contributes to the open innovation literature in three ways. First, our study provides large-scale evidence on the determinants of open innovation for U.S. firms, which heretofore has been limited (West and Boger, 2017). We employ the economically important setting of software development to examine open innovation in a broad sample of firms using the GitHub platform, which has been adopted by 90% of Fortune 100 companies (GitHub, 2023). Second, our study explores the *channels* through which open innovation influences firm performance, adding new insights. Third, as Piller and West (2014) suggest, much of the open innovation literature focuses

[9] In addition to these published sources, in concurrent working papers, Hu et al. (2023) find a positive association between external contributions to a firm's open-source software projects and Tobin's Q, while Yang (2023) finds a positive market response to GitHub code updates for software firms. Our study is unique from this concurrent work, as we focus on the determinants of open innovation for a broad sample of U.S. firms as well as the channels through which open innovation affects firm performance.

on organizations as the key actors (e.g., strategic innovation alliances between firms or firms' purchases of intellectual property from other organizations). In our test of the channels of value creation, we examine the impact of individual developers' contributions to GitHub postings. Our setting thus facilitates knowledge flows between the firm as an *organization* and external *individuals* (i.e., GitHub contributors). As such, we provide evidence on firms' interactions with individual inventors (i.e., developers), which Glaeser and Lang (2024) identify as an emerging innovation research opportunity.

2.3 Innovation Disclosure

While the pursuit of open innovation is a strategic operational choice, it is a strategy that inherently involves public disclosure of invention details, and thus our predictions are also informed by the innovation disclosure literature. Consistent with proprietary cost considerations motivating managers' voluntary innovation disclosure decisions (Verrecchia, 1983; Diamond and Verrecchia, 1991), prior work documents a negative association between empirical proxies for proprietary costs and the provision of voluntary innovation disclosures (Guo et al., 2004; Cao et al., 2018). In our setting, releasing software projects on the GitHub platform can be costly due to the proprietary costs of revealing source code and process-level information. Furthermore, releasing source code under an open license agreement can weaken intellectual property protection, potentially increasing the costs of disclosure. These costs can reduce the expected net benefits from open innovation.

2.4 Empirical Prediction

Given that GitHub posting is a voluntary choice, firms pursuing this strategy plausibly expect the benefits of open innovation to outweigh the costs ex ante. Therefore, we predict that GitHub usage generates value for adopting firms, on average, compared to non-adopting firms. While both adopters and non-adopters make profit maximizing decisions given their strategic position, we expect to observe incremental benefits accruing to open software development firms if open innovation represents an investment with positive net present value (NPV) for adopters but is not NPV positive for non-adopting firms given their current capabilities. As an example, some firms can benefit from advanced machine learning techniques given they have access to specialized data sources, such as large amounts of historical customer data, while other firms may not have previously developed these data sources. Thus, firms with large volumes of customer data can realize greater returns from machine learning whereas other firms cannot, leading to higher predicted future earnings for machine learning firms relative to non-machine learning firms, even though both sets of firms are profit maximizing. Similarly, firms that adopt open innovation are likely better positioned to capitalize on this novel strategy and the profitable investment it provides than firms who choose not to pursue open innovation given their strategic position.

As for *how* releasing information on GitHub benefits firms, we anticipate that the benefits flow through three non-mutually exclusive channels: 1) directly improving the firm’s underlying software projects through crowdsourced software development, 2) facilitating widespread adoption of the firm’s technologies, and 3) providing a source of managerial learning that improves the firm’s future project selection decisions.

First, crowdsourcing is defined as “an entity taking a function once performed internally and outsourcing it to an undefined, large network of individuals in the form of an open call” (Jame et al., 2016). Afuah and Tucci (2012) suggest that crowdsourcing can improve the efficiency and effectiveness of firms’ problem solving. The key components of successful crowdsourcing are a firm-specific task that needs to be completed, a community that is willing and able to perform the task, an online environment that facilitates the work and enables the community to interact with the firm, and a mutual benefit for the firm and the community (Brabham, 2013; Jame et al., 2016). As a collaborative software development platform, GitHub embodies these components, which we expect will enable firms to harness the wisdom of the crowd. As a result, we expect repositories that receive more community engagement to lead to more pronounced value creation for firms.

Second, publicly posting GitHub projects can lead to more widespread adoption of the disclosing firms' software (Bonaccorsi and Rossi, 2003). GitHub increases dissemination, spreading a firm's technologies to a wider user base, allowing others to use and build on the disclosed work (e.g., "innovation diffusion"; Schumpter, 1934). This phenomenon has similarities to firms disclosing their patents to a Standard Setting Organization to ensure a future stream of royalty revenue (Oh et al., 2024). Widespread adoption can benefit the disclosing firm by expanding the number of products and services that rely on the software or by facilitating the sale of consulting and installation services. Thus, we anticipate GitHub activity will lead to higher value creation for firms with more widely disseminated or highly impactful GitHub projects.

Third, we expect GitHub activity to provide a source of managerial learning that improves firms' future project selection decisions. An emerging literature suggests that managers can enhance their learning and improve firm-level investment decisions by soliciting market feedback to voluntary disclosures (Jayaraman and Wu, 2020) and by interacting with informed external parties (Guo and Zhong, 2023). By posting software repositories on GitHub, managers receive feedback on these projects from external developers, including information about the projects' technical potential and the level of public interest. This feedback could improve managers' information set, leading to better project selection and resource allocation decisions (Roychowdhury et al., 2019). Thus, we expect firms that are active on GitHub to generate greater returns for a given level of innovative investment.

However, the prediction that GitHub adopting firms will have greater future earnings is not without tension. By publicly posting source code, managers can facilitate competitors' use of the firm's investment. As previously hypothesized, this widespread adoption could accrue benefits to the GitHub adopting firm but could also represent a proprietary cost that reduces any net benefits. Additionally, posting process-level source code could leave the firm vulnerable to cybersecurity threats.[10] Ultimately, while managers may anticipate the benefits to outweigh the costs when engaging in an open innovation strategy, the realization of outcomes ex post could be different.

3 RESEARCH DESIGN

3.1 Data and Sample

Using the GitHub API, we obtain a list of all organization accounts on the platform. We then match publicly traded firms to GitHub organizations during the years 2008 (when GitHub was launched) to 2021, matching first based on the company website and then by firm name. We observe very little adoption of GitHub during the first two years of the platform's existence; however, GitHub adoption has increased significantly since 2010. Using the list of publicly traded firms on GitHub, we obtain additional data for our analyses, including information on repositories, contributors, and forks. We provide further details regarding the collection of GitHub data in Appendix A. We obtain the remaining data necessary for our study from Compustat, CRSP, I/B/E/S, TAQ, EDGAR 10-K filings, and patent data from Kogan et al. (2017) updated through 2022. Because our analyses include variables that require several years of observations (e.g., five years of future earnings), the sample period used in our primary tests is 2010 to 2017.

In Table 1 Panels A and B, we compare firm-years with GitHub accounts to all firms on Compustat by Fama-French 12 industry groups. We do so for both the full sample—with all available years (2008-2021)—and the main sample with non-missing data necessary for our empirical analyses (2010-2017). We find that a broad cross section of approximately 1,597 unique firms—representing all major industries—adopt GitHub through 2021. Further, we find that the representation across industries, and relative to the Compustat population, is generally similar for both sample periods. Moreover, GitHub-adopting firms add approximately 15 new repositories on

[10] Cybersecurity news reports that "GitHub has revealed that over 39 million secrets were leaked across its platform in 2024 alone… The exposed secrets include API keys, credentials, tokens, and other sensitive authentication data that could give attackers unauthorized access to critical systems and services." https://cybersecuritynews.com/39m-secret-api-keys-credentials-leaked-from-github/. See further discussion in Section 4.5.

average each year, with firms in the business equipment industry posting approximately 19 new repositories each year. We provide additional descriptive information in Section 4.1.

3.2 Empirical Models

3.2.1 Determinants of GitHub Activity

Given the novel nature of our dataset, we first examine the firm characteristics associated with GitHub usage. We estimate the following model:

$$NewGitHub_{i,t} = \alpha_0 + \alpha_{1-5} Innovation_{i,t} + \alpha_{6-12} Investment_{i,t} + \alpha_{13-18} InfoEnviron_{i,t} + IndFE + YrFE \tag{1}$$

In the above model, *i* indexes the firm and *t* indexes year, and we estimate three main categories of determinants: 1) non-GitHub innovative activities (*Innovation*), 2) investment strategy and profitability (*Investment*), and 3) the information environment (*InfoEnviron*). Our first group of determinants is non-GitHub innovative activity that includes empirical proxies for innovations created in-house (*lnSoftwarePatents*, *lnNonSoftwarePatents*, *RD*, and *TradeSecret*) or acquired (*MA*). If open innovation is a complement to a firm's traditional pursuit of innovation, we expect a positive association between the non-GitHub innovation variables and GitHub activity. If open innovation substitutes for another type of innovative activity, we expect a negative association.

Our second overarching group of determinants relates to a firm's investment strategy and profitability. This includes capital expenditures (*CAPEX*), SG&A expenditures (*SGA*), advertising expenditures (*ADV*), adjusted net income during the year (*AdjNI*), the change in adjusted net income from the prior year (*ChAdjNI*), leverage (*Lev*), and implied growth prospects (book-to-market ratio, *BM*). A positive association between the investment and profitability proxies and GitHub usage would suggest managers anticipate incremental benefits from GitHub usage that complement existing investment strategies. A negative association between investment and profitability measures and GitHub usage could indicate that firms with strategies that are underperforming turn to GitHub as an alternative.

Our third group of determinants relates to a firm's information environment. This includes firm size (*MVE*), analyst following (*AnalystFoll*), institutional ownership (*InstOwn*), firm age (*Age*), industry concentration (*HHI*), and financial reporting transparency (*MGMTFore*). In general, if firms with stronger information environments also pursue greater transparency in their innovations, we would expect a positive association between these proxies and GitHub usage. Formal variable definitions are included in Appendix C. Due to the observed increase in GitHub adoption over time and across industries, we also include fiscal year and industry (Fama-French 48) fixed effects in our determinants model.

3.2.2 Open Innovation and Value Creation

Our first test seeks to determine whether open innovation is positively associated with value creation. We use the following OLS regression model:

$$AdjNI[+1,+5]_{i,t} = \alpha_0 + \alpha_1 NewGitHub_{i,t} + \alpha_{2-6} Innovation_{i,t} + \alpha_{7-13} Investment_{i,t} + \alpha_{14-19} InfoEnviron_{i,t} + IndFE + YrFE \quad (2)$$

In the above model, *i* indexes the firm and *t* indexes year. The dependent variable, *AdjNI [+1, +5]*, is the firm's net income before R&D, advertising, and depreciation aggregated from fiscal year *t+1* through fiscal year *t+5*, scaled by total assets at the end of year *t*, following Curtis et al. (2020).[11] We use five-year future earnings to allow sufficient time for GitHub activity to manifest in earnings since GitHub projects in our sample remain active for approximately 4.14 years. Further, this five-year window is consistent with the five-year amortization period of software used for tax purposes and the commonly two- to five-year period used for financial accounting. In this model, we are interested in the coefficient on *NewGitHub*, which is the log of one plus the number of new, public repositories released on GitHub by firm *i* during fiscal year *t.* The coefficient on *NewGitHub* reflects whether an open innovation strategy is associated with future earnings.

[11] We follow Curtis et al. (2020) by adding back R&D, advertising, and depreciation to net income. This adjustment is important, as we find in our determinants analysis that R&D and advertising are significantly positively associated with GitHub adoption (Table 3). Thus, including these expenditures in our dependent variable could lead to a mechanical relation between unadjusted earnings and GitHub activity.

We include several control variables and fixed effects in our models to better isolate the relation between open innovation and future realized earnings. Specifically, we control for several characteristics of the firm that could be correlated with open innovation and future earnings. As more innovative firms tend to have higher future realized earnings (Lev and Sougiannis 1996), and such innovation activity may be correlated with GitHub adoption, we control for the firms' innovative activity as proxied by the number of software patent filings (*lnSoftwarePatents*), non-software patent filings (*lnNonSoftwarePatents*), R&D expenditures (*RD*), trade secrecy (*TradeSecret*) and M&A activity (*MA*). Another purpose of including *lnSoftwarePatents* in our model is to control for a potential concern that firms with more software development activity have both more projects to post on GitHub and higher future earnings simply because of the importance of software to the firm, rather than GitHub usage per se. By controlling for firms' software patents, we more precisely account for the firm's level of software development activity.[12]

We also include as controls the determinants described in Section 3.2.1 to account for changes in the firm's investment strategy over time and other determinants of firm profitability, following Curtis et al. (2020). Similarly, we include industry fixed effects (Fama-French 48) to control for time-invariant, unobserved heterogeneity across industries and allow us to exploit intra-industry variation.[13] We also include year fixed effects to ensure that our results generalize over time. Moreover, we winsorize all continuous variables at the 1st and 99th percentiles by year, and we cluster standard errors by firm. We define all variables in Appendix C.

[12] An ideal experiment would hold constant the quantity and quality of firms' software development activities and vary only the public posting of source code. Controlling for software patents helps mitigate the concern that firms with a heightened focus on software development (and hence, more activity on GitHub) might have greater software development activities (i.e., higher quantity) and attract higher quality developers (i.e., higher quality), which could lead to improved firm performance. Furthermore, we perform two additional analyses (untabulated) to address the potential non-random assignment of developer talent contributing to our results. First, our results are robust if we examine a subsample of firms with software patents and include a control for the talent of software patent inventors. We proxy for the talent of software patent inventors using the average experience of inventors and the number of co-inventors on each software patent, as well as the average number of forward citations per software patent. Moreover, our inferences remain unchanged if we control for the mention of "chief information officer" or "CIO" in the 10-K as a proxy for the importance/talent of developers.

[13] We favor industry fixed effects rather than firm fixed effects given that our research question examines an open innovation paradigm, which is a firm strategic choice. Due to the persistent nature of this strategy choice with effects that manifest over time, exploiting cross-sectional variation in GitHub adoption best addresses our research question. As noted in Breuer and deHaan (2023), within-firm variation is not necessarily superior to cross-sectional variation. We also examine a modified control function method to adjust for selection effects and include a specification with firm fixed effects as further discussed in Section 5.2.2.

4 RESULTS

4.1 Descriptive Statistics

Figure 1 presents a graph of GitHub adoption over time, starting in 2008 when the platform was released. By 2021, approximately 22% of publicly traded firms have an active GitHub account. This rapid increase illustrates the increasing importance of software collaboration and open-source software development over time. At the repository level, we descriptively examine the level of participation from the developer community in Table 1 Panel C for the 2010-2017 sample period. We find that a firm-year has an average of 21.6 accepted collaborations from any source across all repositories. GitHub also categorizes the source of the collaboration as a member, collaborator, contributor, or none. A member is an individual who is part of the organization; a collaborator is an individual who is not part of the organization but has defined access to the repository; a contributor is an external person who has made multiple contributions to a repository; the "none" group is uncategorized. We find that most collaborations originate from the "contributor" category, with over 14 accepted coding suggestions per repository, on average. Moreover, we find that the average repository at the firm-year level has 27 forks, which are developers' copies of posted repositories. Taken together, the descriptive statistics show that GitHub repositories typically have a significant amount of community participation.

Table 2 presents the descriptive statistics for the variables used in our regression analyses. Firms in our sample are profitable on average (based on five-year aggregated income) and have an adjusted, scaled net income of 0.57. Just over five percent of the observations in our sample (including both GitHub adopting and control firms) have at least one new repository posted in a firm-year (untabulated), while the average number of new repositories across the entire sample is 0.39 (*NewGitHub Count*). We find that the firms in our sample have average annual expenditures of R&D equal to approximately 3.6 percent of assets, CAPEX expenditures equal to 4.0 percent of assets, and book-to-market ratios of 0.62.

4.2 Determinants

Table 3 reports our determinants results. The results are presented with and without fixed effects (industry and year) in columns 1 and 2, respectively. In column 2, we find that generally more innovative firms, as proxied by patents (both software patents and non-software patents), R&D expenditures, and trade secrecy, are more likely to use GitHub. These results suggest that firms use GitHub as a complement to, rather than a substitute for, their other innovative activities. This finding is consistent with the observation that firms pursue the development of both open-source and proprietary software (Lin and Rai, 2024).

We also find that advertising (*ADV*) and firm profitability (*AdjNI*) are positively associated with GitHub activity, suggesting that profitable firms that invest in product and reputation building activities are active on the platform. Furthermore, firm size (*lnMVE*), analyst following (*lnAnalystFoll*) and industry concentration (*HHI*) are positively associated with GitHub activity, consistent with an open innovation strategy being a complement to a strong information environment. Finally, our evidence suggests that younger firms are more likely to use GitHub, which is consistent with greater technological focus among these firms. Overall, our determinants analysis contributes to the open innovation literature by systematically documenting in a broad sample of firms the factors associated with adopting an open innovation strategy, using software development as a setting.

4.3 Open Innovation and Value Creation

Table 4 presents our main result examining the association between open innovation and future earnings. We estimate our main analysis in three ways: (1) without controls or fixed effects (column 1) (Whited et al., 2022), (2) with controls, but no fixed effects (column 2) (Jennings et al., 2023), and (3) with controls and fixed effects (column 3). In each column, we find a positive and significant coefficient on *NewGitHub*, which suggests GitHub repository postings are associated with greater future earnings. Importantly, this finding is consistent even when we control for the variables included in our determinants model. In other words, after explicitly controlling for observable differences between GitHub-adopting firms and non-adopting firms, we find a significantly positive association between GitHub activity and value creation. In terms of economic significance, we find that a one standard deviation increase in *NewGitHub* is associated with a 6.85% increase in earnings relative to the unconditional mean value in our sample period.

4.4 Cross-sectional Results

We next examine cross-sectional variation in GitHub activity to better understand the mechanisms through which open innovation, via GitHub, is positively associated with future earnings. We identify three potential channels: (1) crowdsourced engagement, (2) widespread adoption, and (3) managerial learning.

4.4.1 Crowdsourced Engagement

First, we examine the association between the extent of crowdsourced engagement and future earnings by splitting GitHub repositories on the number of accepted external contributions. Specifically, *HighCollab* is the log of one plus the number of new public repositories of firm *i* in year *t* with above median accepted contributions by external "contributors" or "collaborators" by year. We focus on contributions by "contributors" and "collaborators" because we expect these contributions to be the highest quality external contributions. *LowCollab* is the log of one plus the number of repositories with at or below-median accepted contributions by "contributors" and "collaborators". We expect the positive association between GitHub usage and future earnings to be stronger for repositories with more collaboration. Table 5, Panel A presents our results. We find that *HighCollab* is significantly positively associated with future earnings individually (column 1) and when combined with *LowCollab* (column 3). However, *LowCollab* is significantly positively associated with future earnings only when *HighCollab* is not included in the model (column 2). The difference between *HighCollab* and *LowCollab* is statistically significant (column 3). These results suggest the positive association between future earnings and new GitHub repositories is increasing in the amount of crowdsourced engagement each repository receives.

4.4.2 Widespread Adoption

Next, we examine the association between how widely a firm’s GitHub repositories are adopted and future earnings by splitting GitHub repositories based on the level of external adoption. *HighAdopt* is the log of one plus the number of new GitHub repositories with above-median forks by year. *LowAdopt* is the log of one plus the number of new GitHub repositories during the year with at or below-median forks. Forks are the number of times a repository is copied. If the benefits of widespread adoption exceed any proprietary costs of facilitating competitors’ use of the technology, we expect the association between GitHub activity and future earnings to be stronger for firms whose repositories have more widespread adoption. Table 5, Panel B presents our results. We find that *HighAdopt* is significantly positively associated with future earnings individually (column 1) and when combined with *LowAdopt* (column 3). However, *LowAdopt* is insignificantly associated with future earnings in the combined model (column 3). In column 3, we find a statistically significant difference between *HighAdopt* and *LowAdopt*. These results suggest that the positive association between future earnings and new GitHub repositories is increasing in the extent to which a firm’s repositories are widely adopted.

4.4.3 Managerial Learning

Finally, we examine whether open innovation can improve R&D investment decisions via managerial learning. If managers learn from open-sourced feedback and improve their project selection and continuation decisions, we expect to observe greater returns to investments in innovation. We proxy for returns to investments in innovation by examining the association between R&D expenditures and future earnings. If GitHub usage improves investment decisions, we expect a significantly positive coefficient for the interaction of *NewGitHub* and *RD*. Table 5, Panel C presents these results. We include results with all controls (column 1), controls and fixed effects (column 2), and a fully interacted model with controls and fixed effects (column 3). Across all columns we find that the interaction between *NewGitHub x RD* is positively associated with future earnings, suggesting that the returns to R&D investments are increasing in the extent to which firms utilize GitHub.[14] Taken together, the results of our cross-sectional tests raise the bar for alternative explanations, such as selection effects, to fully explain our results and provide novel evidence on the channels through which open innovation can benefit firms.

4.5 Costs of GitHub Adoption

In this section, we investigate the potential downside costs of GitHub adoption. First, anecdotal evidence suggests that GitHub adoption could expose firms to cybersecurity breaches.[15] We obtain data on cybersecurity breaches for firms in our sample to examine this potential cost. We form three new dependent variables, *Cyber Breach [+1]*, *Cyber Breach [+1, +3]*, and *Cyber Breach [+1, +5]*, which are indicator variables set equal to one if the firm experiences a data breach in year t+1, between years t+1 and t+3, or between years t+1 and t+5, respectively, and zero otherwise. We then regress these variables on a firm's GitHub activity in year t. Because some firms may be subject to cyber security risks more generally due to their industry or business model, we address this possibility by including an additional control variable, *Cybersecurity_10K*, which is an indicator if the firm mentions cybersecurity-related phrases in the risk factor section of the 10-K, and zero otherwise (see Appendix C for variable definitions). Table 6 reports these results. Across each time horizon, we find a significantly positive association between GitHub activity and the likelihood of future data breaches. Overall, these results are consistent with the notion that GitHub adoption may expose firms to cybersecurity risks, reducing the potential net benefits.

[14] In untabulated analysis, we re-estimate our results altering the treatment of missing R&D values. Specifically, we 1) exclude all firm-years with missing R&D, 2) exclude all firm-years with non-positive R&D, 3) include an indicator for firm-years with missing R&D, 4) add capitalized software to R&D, 5) replace missing R&D values with the industry-year average, and 6) include an indicator for firm-years with missing R&D and a patent filing in the same year (Koh and Reeb, 2015). In each of these specifications, our inferences are unchanged.

[15] For example, software security experts report sensitive data leaks, such as access tokens and signing keys, even affecting large technology companies. See https://www.techtarget.com/searchitoperations/news/366621078/GitHub-Actions-supply-chain-attack-spotlights-CI-CD-risks.

5 ADDITIONAL ANALYSES AND ROBUSTNESS

5.1 Short Window Returns

Our primary analyses investigate the relationship between open innovation and future earnings. However, earnings represent the realization of many different investments, which inherently limits our ability to attribute earnings changes solely to GitHub usage. As an alternative proxy for value creation at the more granular project level, we examine the short window returns to firms' public GitHub repository posts. Although some investors may be unaware of firms' GitHub activity, anecdotal evidence suggests that sell-side equity analysts monitor firms' GitHub profiles for updates (Bora, 2022). If GitHub activity is associated with positive returns, this finding would be consistent with open innovation creating value for firms.[16]

We employ a standard event-based methodology and estimate the average buy-and-hold three-day return for each firm-repository posting. We utilize the repository creation date, obtained via the GitHub API, to determine the posting date. Importantly, our returns are risk-adjusted based on the firm's corresponding size, book-to-market, and momentum quintile return (Daniel et al., 1997). We perform the returns tests for all years for which we have significant GitHub adoption (2010-2021) and for the sample period used in our future earnings tests (2010-2017).

Table 7 reports our results. In Panel A, we find strong evidence of a positive market response to open innovation activity during the years 2010-2021. Specifically, following GitHub repository postings, firms exhibit positive, statistically significant three-day returns with an economic magnitude of approximately 5.9 basis points (5.1% annualized). During the 2010 to 2017 sample

[16] We note that a positive return could have two components. First, investors may have a positive reaction to the signal about firms' ongoing innovative activities that GitHub project postings may reveal. Second, investors could anticipate benefits from open innovation specifically. Ideally, we would measure only the second component and not the first, but returns-based tests inherently confound the two effects. However, at a minimum, we expect that if investors predict that the costs of open innovation exceed the value of the underlying technology, we would not find a positive return. Ultimately, our returns tests, in connection with the earnings-based tests in our main analyses, help to triangulate the effect of open innovation.

period used in our main tests (Panel B), we find positive and significant returns to GitHub repository postings with an economic magnitude of 7.9 basis points (6.9% annualized).[17]

Finally, we partition the sample into repositories with high versus low adoption (based on copies). We find that the average market response increases to 11.7 basis points (10.3% annualized) across our main sample period for widely adopted repositories, which are those that have copies that are greater than or equal to the sample median. These findings suggest that, at the project level, investors anticipate benefits from GitHub activity. Overall, this evidence corroborates the results from our earnings prediction tests.

5.2 Robustness

5.2.1 Alternative Samples

We perform several robustness analyses to alleviate concerns that our results are attributable solely to selection effects. Our first approach seeks to compare the sample of GitHub-adopting firms to two alternative control samples. While our main tests use all Compustat firms as control firms given the ubiquitous nature of software in the modern economy, our first alternative sample applies a more tailored control sample of firms whose business activities lend themselves to participating in meaningful software development. We identify three indicators of software development activity (i.e., the existence of an active GitHub account, non-zero capitalized software, and non-zero software patent filings) and include only firm-years in our sample if they belong to an SIC 4-digit industry where at least one industry member has at least one indicator of software development in a fiscal year. This control group arguably holds constant the economic conditions that are common to software development and exploits variation in GitHub activity.

[17] In untabulated robustness analyses, we conduct a placebo event date test by analyzing the firms' short-window returns beginning ten days prior to each repository posting. We find insignificant returns for this placebo analysis, which suggests the positive returns we observe following repository postings are unlikely to be due to positive firm momentum. Further, our inferences remain unchanged if we adjust the returns using the overall market value-weighted return (rather than the characteristic adjusted return for each firm).

Our second alternative approach limits the sample to include GitHub-adopting firms only. Specifically, we retain firms that have had a publicly posted repository on GitHub during the sample period and exploit variation in the number of new, public repositories posted, thereby holding constant a firm's choice to adopt the platform. While our main results in Table 4 speak to the extensive margin effect of GitHub adoption, this analysis holds constant the choice to adopt GitHub and provides evidence at the intensive margin.

We present these alternative control sample results in Table 8, Panel A. In column 1, we continue to find a positive and significant coefficient on *NewGitHub* in the subsample of firms with an indication of software development activity. This result suggests that even after we limit our sample to firms most likely to have economic activities calling for software development, *NewGitHub* is still strongly associated with future earnings. Furthermore, our sample size is not significantly reduced using this tailored set of control firms (16,626 observations compared to the full sample of 20,302), corroborating the assertion that software development activity is important to the vast majority of firms in our sample period.

In column 2, we continue to find a positive and significant association between the posting of new GitHub repositories and future earnings in the sample that excludes firms without any GitHub activity during our sample period. Taken together, the evidence from our alternative control samples finds both an extensive and intensive margin effect of GitHub activity, which points to GitHub activity contributing to the association with future earnings rather than being solely attributable to differences between open innovators and other firms.[18]

5.2.2 Modified Control Function

While we control for observable differences between GitHub adopting and other firms, it remains possible that measurement error or unmodeled differences between treated and control

[18] We also find in untabulated analysis that our results are robust to excluding tech firms, suggesting that the findings generalize across non-tech industries and are not driven solely by the largest technology companies (e.g., Apple, NVIDIA, Microsoft, etc.).

firms contribute to the observed effect. To provide additional evidence of the association between *NewGitHub* and future firm performance, we apply the Klein and Vella (2010) modified control function approach. This approach adds a modified control function variable that explicitly controls for the potential endogeneity inherent in the relation between GitHub activity and performance, while not relying on the existence of a narrowly defined instrument (Klein and Vella, 2010).

To apply this approach, we follow Armstrong et al. (2022) by (1) making required assumptions about the source of endogeneity, (2) estimating the endogeneity through the unexplained portion of the first-stage model, and (3) controlling for the endogeneity in the main model. This approach decomposes the error term from equation 2 into the endogenous component of the unobservable factors that drive both innovation activities and future firm performance as well as the exogenous aspect.[19] A potentially important source of endogeneity could be due to GitHub adopting firms having greater current performance and growth potential (i.e., highly profitable, growth firms are more likely to have greater future performance even absent GitHub activity). Therefore, we sort firms into terciles by year based on current performance (*AdjNI*) and growth potential (*BM*). After sorting firms on both dimensions, we predict the endogenous component of the relation by regressing *NewGitHub* on the variables in our determinants model to calculate the unexpected level of GitHub activity. After estimating this first-stage model, we then estimate the endogenous relation between future firm performance and current GitHub activity by examining the variation in unexplained *NewGitHub* and *AdjNI [+1, +5]*. We iteratively estimate the variation in *NewGitHub* and *AdjNI [+1, +5]* until the endogenous variable converges.

Table 8, Panel B presents the results of equation 2 using the modified control function approach. We present our results using industry and year fixed effects (column 1) and with firm and year fixed effects (column 2). In each specification, we follow Armstrong et al. (2022) and

[19] We thank Chris Armstrong, Allison Nicoletti, and Frank Zhou for sharing their code to perform the modified control function test. See Armstrong et al. (2022) and Timmermans (2024) for a more detailed explanation of the modified control function approach.

estimate bootstrapped standard errors (using 1,000 bootstrapped samples) to address concerns that influential observations may artificially inflate statistical significance. We also report 90% confidence intervals for all variables of interest. We continue to find a positive and significant association between posting new repositories on GitHub and future five-year ahead performance.

We also note that the coefficient for ρ is negative and statistically significant in column 2, suggesting GitHub activity is at least partially endogenously related to future firm performance. However, after explicitly controlling for this endogeneity and observable firm differences, as well as including firm fixed effects, the positive association between GitHub activity and future performance persists. Overall, this finding suggests that while selection effects contribute to the relation, they are unlikely to fully explain the results.[20]

5.2.3 Alternative Measurements and Specifications

We next address additional concerns relating to research design using alternative measurement techniques, fixed effect structures, and outlier treatment. Our first set of robustness tests examine alternative measurement approaches for GitHub activity and future performance. Our main analysis employs the logged number of new public GitHub repositories as the regressor of interest. We re-estimate our model by replacing our independent variable of interest with (1) the inverse hyperbolic sine of the count of new repositories, (2) the unlogged count of the number of new repositories, (3) an indicator for new repositories, and (4) an alternative measure of GitHub activity that represents the number of a firm's GitHub repositories that continue to receive external participation or updates during year *t*. We report these results in Table 8, Panel C. Using each of these alternative GitHub activity measurement approaches, we find a significantly positive

[20] We view our study as providing important early evidence on the choice to adopt an open innovation strategy as well as highlighting potential benefits and costs from pursuing non-excludable innovations at the corporate level. While we take reasonable steps such as controlling for observable differences between adopting and non-adopting GitHub firms, triangulating inferences via cross-sectional tests that corroborate the mechanism, and examining the robustness of our results to numerous alternatives, we acknowledge that our tests are association-based. We welcome and encourage future research that might tighten causal inferences in this setting.

association between GitHub usage and future earnings, which suggests that our results are not attributable to a specific type of measurement of GitHub activity.

Next, we examine alternative measures of future performance. Our main analysis aggregates future adjusted net income over the next five years following Curtis et al. (2020). We continue to follow Curtis et al. (2020) and change the dependent variable to reflect aggregated future sales, future earnings over the next three years, and future earnings in the next year, each scaled by assets. Moreover, we use an alternative future earnings model following Ciftci and Cready (2011). Table 8, Panel D reports these results. We continue to find a positive and significant association between GitHub usage and these alternative measures of future performance. Lastly, we also modify equation 2 to separately regress *AdjNI* in year t+1, t+2, t+3, t+4, t+5 on *NewGitHub* and controls to avoid any bias that results from employing a multi-year dependent variable. We continue to find that the association between future earnings and GitHub usage continues to be significantly positive in all time horizons (untabulated).

Our final robustness approach alters the specific fixed effect structures and methods for dealing with outliers. Specifically, we employ alternative classifications of industry fixed effects (i.e., Fama-French 12, two-digit SIC code, and three-digit SIC code) as well as industry by year fixed effects. Moreover, we re-estimate our main results after omitting outliers with Cook’s distance values greater than 4/N or studentized residuals values greater than 2 as recommended by Leone et al. (2019). We continue to find consistent results (untabulated).

6 CONCLUSION

We examine firms’ strategic choice to pursue an open innovation strategy and the potential benefits this approach can yield for adopting firms. Specifically, we examine firms’ disclosure of software source code on the collaborative software development platform GitHub as a type of open innovation. We find that firms use open innovation as a complement to traditional sources of intellectual property protection such as patenting and trade secrecy. Our findings also suggest that publicly disclosing innovation is linked to value creation for adopting firms, as evidenced by a positive association between GitHub activity and greater future earnings as well as positive risk-adjusted market returns to project postings. Cross-sectional results suggest that the positive association between open innovation and future earnings is increasing in the extent to which GitHub repositories receive high-quality crowdsourced suggestions and are more widely adopted. Furthermore, we find that the returns on innovative investment are greater for firms active on GitHub, consistent with the potential for open innovation to improve innovation investment decisions. We also find some evidence suggesting that open innovation can expose firms to cybersecurity risks.

Our results should be interpreted with caution as they represent first steps toward understanding the phenomenon of open innovation on a large scale. While our results are consistent using alternative design choices, vary as expected in the cross-section, and are robust to numerous additional tests, we emphasize that we exploit conditional correlations rather than exogenous variation. The choice inherent in a firm's pursuit of an open innovation strategy leaves open the possibility that omitted factors or mismeasurement contribute to the observed association. Accordingly, our results can be interpreted as contributing to the "inventory of potential economic outcomes" for firms that voluntarily adopt GitHub (Leuz and Wysocki 2016), and the benefits we document may not manifest among firms that are compelled to pursue open innovation. Furthermore, while we take empirical steps to mitigate the *firm-level* selection effects of choosing to post source code on GitHub, it is also important to note that there are potential *project-level* selection effects. Thus, even within a GitHub-adopting firm, managers choose which projects to keep secret and which projects to publicly post. While we embrace the voluntary nature of GitHub activity and document its complementary nature with firms' other innovative pursuits, we highlight this point to help readers interpret our results. Future researchers can refine our data in ways that could improve project-level inferences, such as developing a classification system for repository type or mapping GitHub projects to products, patents, or segment data.

Our new dataset of public firm usage of open innovation via GitHub can be of use to future researchers in multiple disciplines. For example, researchers could investigate additional costs or benefits of GitHub disclosure, including benefits to complementary proprietary business lines, deterring competition, recruiting developers, or advertising the firm prior to raising capital or in the market for corporate control (Bourveau et al., 2022). We leave these extensions of this research agenda to future research.

References

Afuah, A., Tucci, C. L. 2012. Crowdsourcing as a solution to distant search. *Academy of Management Review* 37, 355-375.

Armstrong, C., Glaeser, S., Park, S., Timmermans, O. 2024. The assignment of Intellectual Property Rights and Innovation (Unpublished Working Paper).

Armstrong, C. Nicoletti, A., Zhou, F. S. 2022. Executive stock options and systemic risk. *Journal of Financial Economics* 146, 256-276.

Baruffaldi, S. H., Simeth, M., Wehrheim, D., 2024. Asymmetric Information and R&D Disclosure: Evidence from Scientific Publications. *Management Science* 70, 1052-1069.

Bigliardi, B., Ferraro, G., Filippelli, S., Galati, F. 2020. The influence of open innovation on firm performance. *International Journal of Engineering Business Management* 12, 1-14.

Blankespoor, E., Miller, G. S., White, H. D. 2014. The role of dissemination in market liquidity: Evidence from firms' use of Twitter™. *The Accounting Review* 89, 79-112.

Blankespoor, E., deHaan, E., Marinovic, I. 2020. Disclosure processing costs, investors' information choice, and equity market outcomes: A review. *Journal of Accounting and Economics* 70 , 1-46.

Bonaccorsi, A., Rossi, C. 2003. Why Open Source Software Can Succeed. *Research Policy* 32, 1243–1258.

Bora, P. 2022. Elastic Forks Out a Solid Consistent Quarter. North America Equity Research. J.P. Morgan.

Boudreau, K. J. 2010. Open Platform Strategies and Innovation: Granting Access vs. Devolving Control. *Management Science* 56, 1849–1872.

Boudreau, K. J., Lacetera, N., Lakhani, K. R. 2011. Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis. *Management Science* 57, 843–863.

Bourveau, T., De George, E. T., Ellahie, A., Macciocchi, D. 2022. The Role of Disclosure and Information Intermediaries in an Unregulated Capital Market: Evidence from Initial Coin Offerings. *Journal of Accounting Research* 60, 129–167.

Brabham, D. C. 2013. Crowdsourcing. The MIT Press Essential Knowledge Series. Cambridge, Massachusetts ; London, England: The MIT Press.

Breuer, M., deHaan, E. 2024. Using and Interpreting Fixed Effects Models. *Journal of Accounting Research* 62 (4): 1183-1226.

Cao, S. S., Ma, G., Tucker, J. W., Wan, C. 2018. Technological Peer Pressure and Product Disclosure. *The Accounting Review* 93 95–126.

Cao, S. S., Cheng, S., Tucker, J. W., Wan, C. 2018. Technological peer pressure and skill specificity of job postings. *Contemporary Accounting Research* 40 (3): 2106–2139.

Chesbrough, H. 2004. Managing Open Innovation. *Research-Technology Management* 47, 23–26.

Chesbrough, H. 2006. Open Innovation: The New Imperative for Creating and Profiting from Technology. Harvard Business Review Press.

Chu, J., He, Y., Hui, K. W., Lehavy, R. 2025. New Product Announcements, Innovation Disclosure, and Future Firm Performance. *Review of Accounting Studies* 30: 352–383.

Ciftci, M., Cready, W. M. 2011. Scale Effects of R&D as Reflected in Earnings and Returns. *Journal of Accounting and Economics* 52, 62–80.

Curtis, A., McVay, S., Toynbee, S. 2020. The Changing Implications of Research and Development Expenditures for Future Profitability. *Review of Accounting Studies* 25, 405–37.

Dahlander, L., Gann, D. M. 2010. How Open Is Innovation? *Research Policy* 39, 699–709.

Daniel, K., Grinblatt, M., Titman, S., Wermers, R. 1997. Measuring Mutual Fund Performance with Characteristic-Based Benchmarks. *The Journal of Finance* 52, 1035–1058.

Diamond, D. W., Verrecchia, R. E. 1991. Disclosure, Liquidity, and the Cost of Capital. *The Journal of Finance* 46, 1325–1359.

Dyer, T. A., Glaeser, S., Lang, M. H., Sprecher, C. 2024. The Effect of Patent Disclosure Quality on Innovation. *Journal of Accounting and Economics* 77, 101647.

Gartner. 2024. Retrieved November 11, 2024 from https://www.gartner.com/en/newsroom/press-releases/2024-10-23-gartner-forecasts-worldwide-it-spending-to-grow-nine-point-three-percent-in-2025

GitHub. 2023. Retrieved November 27, 2023, from github.com/about.

Glaeser, S., Guay, W. 2017. Identification and generalizability in accounting research: A discussion of Christensen, Floyd, Liu, and Maffett 2017. *Journal of Accounting and Economics* 64, 305-312.

Glaeser, S. 2018. The Effects of Proprietary Information on Corporate Disclosure and Transparency: Evidence from Trade Secrets. *Journal of Accounting and Economics* 66, 163–193.

Glaeser, S., Landsman, W. R. 2021. Deterrent Disclosure. The Accounting Review 96, 291–315.

Glaeser, S., Lang, M. H. 2024. A Review of the Accounting Literature on Innovation. *Journal of Accounting and Economics*. 78, 101720.

Glaeser, S., Michels, J., Verrecchia, R. E. 2020. Discretionary Disclosure and Manager Horizon: Evidence from Patenting. *Review of Accounting Studies* 25, 597–635.

Greco, M., Grimaldi, M., Cricelli, L. 2016. An analysis of the open innovation effect on firm performance. *European Management Journal* 34, 501-516.

Guo, R., Lev, B., Zhou, N. 2004. Competitive Costs of Disclosure by Biotech IPOs. *Journal of Accounting Research* 42, 319–355.

Guo, R., Zhong, R. 2023. Do Managers Learn from Analysts about Investing? Evidence from Internal Capital Allocation. *The Accounting Review* 98, 215–246.

Hu, J., Hu, D., Yang, X. 2023. Can firms improve performance through external contributions to their open-source software projects? (Unpublished Working Paper).

Jame, R., Johnston, R., Markov, S., Wolfe, M. C. 2016. The Value of Crowdsourced Earnings Forecasts. *Journal of Accounting Research* 54, 1077–1110.

Jayaraman, S., Wu, J. S. 2020. Should I stay or should I grow? Using voluntary disclosure to elicit market feedback. *Review of Financial Studies* 33, 3854-3888.

Jennings, J., Kim, J. M., Lee, J., Taylor, D. 2023. Measurement Error, Fixed Effects, and False Positives in Accounting Research. *Review of Accounting Studies* 29, 959-995.

Johnson, J. P. 2014. Defensive Publishing by a Leading Firm. *Information Economics and Policy* 28, 15–27.

Jones, D. A. 2007. Voluntary Disclosure in R&D-Intensive Industries. *Contemporary Accounting Research* 24, 489–522.

Kim, J., Valentine, K. 2021. The Innovation Consequences of Mandatory Patent Disclosures. *Journal of Accounting and Economics* 71, 101381.

Klein, R., Vella, F. 2010. Estimating a Class of Triangular Simultaneous Equations Models without Exclusion Restrictions. *Journal of Econometrics* 154 (2): 154–64.

Knudsen, M. P., Mortensen, T. B. 2011. Some immediate–but negative–effects of openness on product development performance. *Technovation* 31, 54-64.

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

Koh, P., Reeb, D. M. 2017. Missing R&D. *Journal of Accounting and Economics* 60, 73–94.

Laforgia, C. 2019. Your Developers Could Be Publicly Disclosing Source Code By Using Third-Party Code Repositories. https://ipwatchdog.com/2019/04/30/developers-publicly-disclosing-source-code-using-third-party-code-repositories/id=108719/

Laursen, K., Salter, A. 2006. Open for Innovation: The Role of Openness in Explaining Innovation Performance among U.K. Manufacturing Firms. *Strategic Management Journal* 27, 131–150.

Leone, A. J., Minutti-Meza, M., Wasley, C. E. 2019. Influential Observations and Inference in Accounting Research. *The Accounting Review* 94, 337–364.

Leuz, C., Wysocki, P. D. 2016. The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestions for Future Research. *Journal of Accounting Research* 54, 525–622.

Lev, B., Sougiannis, T. 1996. The Capitalization, Amortization, and Value-Relevance of R&D. *Journal of Accounting and Economics* 21, 107–138.

Lin, Y., Rai, A. 2024. The Scope of Software Patent Protection in the Digital Age: Evidence from Alice. *Information Systems Research* 35 (2): 657-672.

Lu, Q., Chesbrough, H. 2022. Measuring open innovation practices through topic modelling: Revisiting their impact on firm financial performance. *Technovation* 114, 102434.

Majchrzak, A., Malhotra, A. 2013. Towards an Information Systems Perspective and Research Agenda on Crowdsourcing for Innovation. *Journal of Strategic Information Systems* 22, 257–268.

Merkley, K. J. 2014. Narrative Disclosure and Earnings Performance: Evidence from R&D Disclosures. *The Accounting Review* 89, 725–57.

Oh, J., Yeung, P. E., Zhu, B. 2024. Technology Coopetition and Voluntary Disclosures of Innovation. *The Accounting Review*, 99(6), 351-388.

Piller, F., West, J. 2014. Firms, users, and innovation: An Interactive Model of Coupled Open Innovation. *New Frontiers in Open Innovation* 29, 29-49.

Roychowdhury, S., Shroff, N., Verdi, R. S. 2019. The Effects of Financial Reporting and Disclosure on Corporate Investment: A Review. *Journal of Accounting and Economics* 68, 101246.

Romer, P. M. 1990. Endogenous technological change. *Journal of Political Economy*, 98(5, Part 2), S71-S102.

Scacchi, W., Feller, J., Fitzgerald, B., Hissam, S., Lakhani, K. 2006. Understanding Free/Open Source Software Development Processes. *Software Process: Improvement and Practice* 11, 95–105.

Schumpter, J. 1934. The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle. Harvard Economic Studies.

Timmermans, O. 2024. Cash versus share payouts in relative performance plans. *The Accounting Review* 96: 451–489.

Verrecchia, R. E. 1983. Discretionary Disclosure. *Journal of Accounting and Economics* 5, 179–194.

West, J., Bogers, M. 2017. Open innovation: current status and research opportunities. *Innovation* 19, 43-50.

Whited, R. L., Swanquist, Q. T., Shipman, J. E., Moon, J. R. 2022. Out of Control: The (Over) Use of Controls in Accounting Research. *The Accounting Review* 97, 395–413.

Yang, W. 2023. How Can Open Source Technology Ecosystem Create Value? Evidence from Investors' Reactions to Firms' GitHub Code Releases. (Unpublished Working Paper).

Appendix A: GitHub Data Collection Process

We access the public GitHub Rest API (api.github.com) to download and create our panel of firm-year GitHub data. We begin by downloading all available organization information from the organization endpoint of GitHub's API. The GitHub API provides both historical and point-in-time data. Our data ends on July 15, 2022.[21] We match GitHub organizations to public firms in two distinct ways.

First, we match on firm website, as GitHub allows organizations to include a website as part of their organization profile. This yields a sample of 1,945 unique GitHub-Compustat firm website matches. Second, we fuzzy match on organization name (GitHub organization to Compustat firm name) and require the match to be at least 95%, which yields an additional 2,608 matched organizations between GitHub and Compustat. The additional matching process using fuzzy matching is important because not all public firms list websites in their GitHub profile.[22] These two steps result in a total sample of organization-firm matches of 4,553, which represents 1,994 unique firms. The total number of organization-firm matches is larger than the unique number of firms because each publicly traded firm can have multiple matched GitHub organizations. For example, IBM has multiple public organizations (e.g., "IBM" and "IBM-CLOUD").[23]

After creating our organization-firm match sample, we then download all repositories for our sample of 4,553 organizations. The repository endpoint of the GitHub API contains all the metadata for users' repositories (i.e., date posted, date of last update, etc.). This download process yields a sample of 58,946 total public repositories from our organization-firm matches. From this step, we use the date posted, date of last update, number of stargazers, forks, and originality in our analyses in the paper.

A repository is considered new if the date posted falls during the fiscal year. Moreover, a repository is considered active if it was posted in a previous fiscal year and is still active during the current fiscal year. Forks is the count of the number of GitHub users who have made copies of the repository. For the high adoption variable (*HighAdopt*), we determine the median level of forks for all repositories posted during the year and all repositories above that benchmark are considered high adoption. Because we use point-in-time data for these tests, we calculate the within-year median when forming *HighAdopt*, as doing so addresses the concern that repositories posted later in the sample period have less time to accrue forks. Finally, to identify whether a repository originated with the firm we use the "forked" column which indicates whether the project is forked from another organization.

We then download all collaborations for the repositories in our sample. Collaborations can either be accepted or rejected by the owner of the repository and are grouped into four categories: member, collaborator, contributor, and none. "Member" refers to individuals who are part of the organization; a "collaborator" is not part of the organization but has defined read, write, or edit permissions for the repository; a "contributor" has made multiple external contributions on a given repository; and the remaining types of collaborations are uncategorized (i.e., "none").

[21] We cannot rule out the possibility that some firms that previously had a GitHub account deleted their accounts prior to our point-in-time download. However, we would expect these deleted accounts to bias against finding results as GitHub treated firms would be incorrectly identified as untreated firms.

[22] Walmart, Bank of America, and Waste Management each have GitHub pages that did not match on firm website from the API.

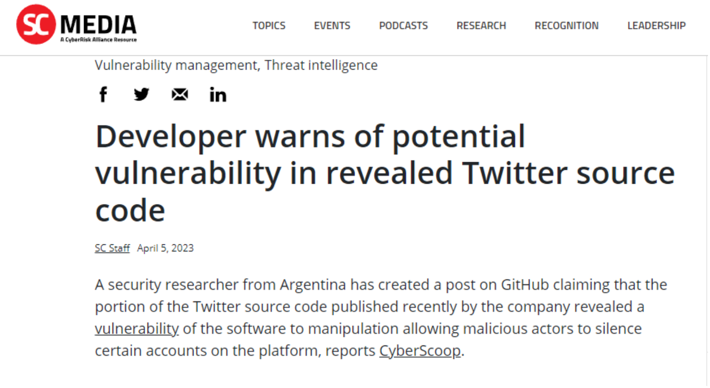
[23] Due to concerns that we incorrectly identify a GitHub organization being linked to a publicly traded firm, we re-estimate our main results and exclude "fuzzy" name matches and find consistent results.

In creating our High Collaborations (*HighCollab*) measure, we seek to identify the amount of high-quality, external contributions to a project. We first eliminate all member collaborations because we are interested in the external contributions. We also eliminate collaborations by "none" because we expect these to be of lower quality. *HighCollab* is the number of repositories with above median accepted contributions that originated from either a "contributor" or "collaborator" for the repository. Again, we employ contributions made by collaborators and contributors because we expect the quality of their coding suggestions to be higher. Similar to our calculation of *HighAdopt*, we determine the median level for all repositories during the year and all repositories above that benchmark are considered high collaboration repositories.

The data we provide to use in future research contains the following: GVKEY, GitHub Organization Name, Link to GitHub Organization API page, Date Posted, Date of Last Update, Stargazers, Forks, Collaboration Members, Collaboration Contributor, Collaboration Collaborator, Collaboration None, and Originality.

Appendix B: Potential Benefits of GitHub Usage

Example One: GitHub Harnesses the Wisdom of the Crowd



Vulnerability management, Threat intelligence

Developer warns of potential vulnerability in revealed Twitter source code

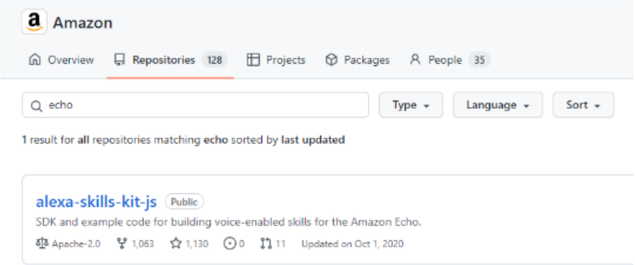
SC Staff April 5, 2023

A security researcher from Argentina has created a post on GitHub claiming that the portion of the Twitter source code published recently by the company revealed a vulnerability of the software to manipulation allowing malicious actors to silence certain accounts on the platform, reports CyberScoop.

Description: Twitter (now known as X) released components of its underlying source code via GitHub on 4/1/2023. Not long after, a third-party developer found a security vulnerability in the code. This highlights an example of how publicly disclosing code on GitHub allows the publisher to harness crowdsourced suggestions and feedback, thereby improving the technology.

Example Two: GitHub Facilitates Widespread Adoption of a Firm’s Products

KidKraft®
Live • Learn • Play



echo

1 result for **all** repositories matching **echo** sorted by **last updated**

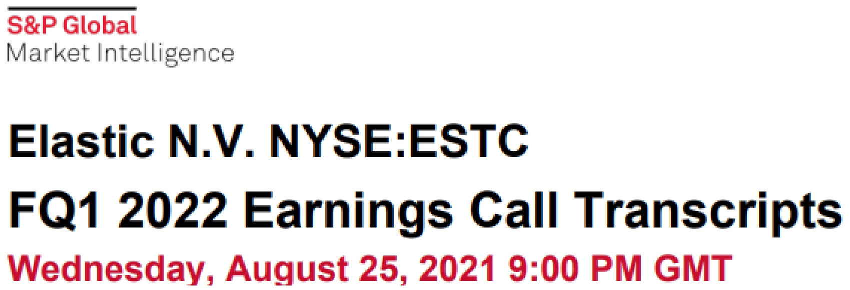
alexa-skills-kit-js Public

SDK and example code for building voice-enabled skills for the Amazon Echo.

Apache-2.0 1,063 1,130 0 11 Updated on Oct 1, 2020

Description: Amazon released code on GitHub enabling third parties to build skills for its Amazon Echo product. This encourages widespread adoption of Amazon’s Alexa-enabled products. For example, KidCraft released a new children’s play kitchen (pictured above) that integrates with Amazon Echo.

Example Three: GitHub Leads to Managerial Learning and Improved Investment Decisions



S&P Global
Market Intelligence

Elastic N.V. NYSE:ESTC
FQ1 2022 Earnings Call Transcripts
Wednesday, August 25, 2021 9:00 PM GMT

*…we’re running ahead fast and innovating for our customer base, and that’s really resonating also with our community. Based on all the metrics that we’re looking at, **downloads**, **engagement on** forums and **GitHub** and others, our users are just running ahead with us and are very excited about all the innovations that we do across all of our 3 solutions and our foundational search platform.*

-Shay Banon, Co-Founder, CEO & Executive Chairman (Elastic N.V.)

Description: In a recent conference call discussion, Shay Banon from Elastic N.V. mentions how the firm monitors GitHub engagement and downloads to inform its innovation-related decisions.

Appendix C: Variable Definitions

Variable	*Description*
ActiveGitHub	is the log of one plus the number of public repositories on GitHub that receive external contribution or updates by firm i during fiscal year t.
AdjNI	is the net income before R&D, advertising and depreciation, scaled by total assets at the end of year t, following Curtis et al. (2020).
AdjNI [+1]	is the AdjNI from fiscal year t+1, scaled by total assets at the end of year t, following Curtis et al. (2020).
AdjNI [+1, +3]	is the AdjNI aggregated from fiscal year t+1 through fiscal year t+3, scaled by total assets at the end of year t (after aggregating), following Curtis et al. (2020).
AdjNI [+1, +5]	is the AdjNI aggregated from fiscal year t+1 through fiscal year t+5, scaled by total assets at the end of year t (after aggregating), following Curtis et al. (2020).
ADV	is advertising expenditures, scaled by total assets at the end of year t, following Curtis et al. (2020).
Age	is the number of years since firm i was listed on CRSP.
AnalystFoll	is the number of analysts who issued earnings per share forecasts for firm i in fiscal year t.
BM	is the book-to-market ratio in fiscal year t.
CAPEX	is the capital expenditures scaled by total assets at the end of year t, following Curtis et al. (2020).
Cash	is the total cash scaled by total assets for firm i year t.
ChAdjNI	is the change in net income before R&D, advertising, and depreciation from fiscal year t−1 to fiscal year t, scaled by total assets at the end of year t, following Curtis et al. (2020).
Cyber Breach [+1]	is an indicator if the firm reports a cyber security breach in year t+1, and zero otherwise.
Cyber Breach [+1, +3]	is an indicator if the firm reports a cyber security breach between years t+1 and t+3, and zero otherwise.
Cyber Breach [+1, +5]	is an indicator if the firm reports a cyber security breach between years t+1 and t+5, and zero otherwise.
CyberSecurity_10K	is an indicator if the firm mentions at least one of the phrases “data breach”, “cybersecurity”, “'cyber security”, “data security breach”, “cyber incident”, “cyber security incident”, “data incident”, or “data security incident” in their risk factor section of the 10-K filing, and zero otherwise.
HHI	is the Herfindahl Index measuring the proportion of total size of firm i in relation to its industry in year t, which we pull from TAQ.
HighAdopt	is the log of one plus the number of new public repositories of firm i in year t with above-median “forks” relative to all other new public repositories in year t.
HighCollab	is the log of one plus the number of new public repositories of firm *i* in year *t* with above-median accepted contributions by “contributors” or “collaborators” relative to all other new public repositories in year t.
ihs_NewGitHub	is the inverse hyperbolic sine of the number of new public repositories issued on GitHub by firm i in year t.
InstOwn	is the percentage of shares owned by institutional investors for firm i in year t.
Lev	is the total current and long-term debt scaled by total assets.
LowAdopt	is the log of one plus the number of new public repositories of firm i in year t with at or below-median “forks” relative to all other new public repositories in year t.

LowCollab	is the log of one plus the number of new public repositories of firm *i* in year *t* with at or below-median accepted contributions by “contributors” or “collaborators” relative to all other new public repositories in year t.
MA	is the total value of M&A deals from CRSP scaled by total assets at the end of year t.
MGMTFore	is an indicator set equal to 1 if firm i issued at least one management earnings forecast during fiscal year t and 0 otherwise.
MVE	is the firm’s market value of equity in fiscal year t.
NewGitHub	is the number of new public repositories issued on GitHub by firm i during fiscal year t. We take the natural log of one plus this variable in the regression analyses.
NewGitHub Count	is the unlogged number of new public repositories issued on GitHub by firm i during fiscal year t.
NonSoftwarePatents	is the number of patents filed by firm i during fiscal year t excluding *SoftwarePatents*.
Op Earn [+1, +5]	is the average operating earnings over five subsequent years from t+1 to t+5 divided by sales revenue in year t. Following Lev and Sougiannis (1996) and Ciftci and Cready (2011), we calculate operating earnings as the sum of operating income before depreciation, advertising, and R&D expenditures.
ρ	is the endogeneity variable that captures the degree to which an unobservable factor is correlated with variation in both the dependent variable and the treatment variable.
Sales [+1, +5]	is the total sales aggregated from fiscal year t+1 through fiscal year t+5, scaled by total assets at the end of year t (after aggregating), following Curtis et al. (2020).
SGA	is selling, general, and administrative expenditures, excluding R&D and advertising, scaled by total assets at the end of year t, following Curtis et al. (2020).
SoftwarePatents	is the number of patents filed by firm i during fiscal year t related to software.
TradeSecret	is an indicator set equal to 1 if the firms’ 10-K filing mentions “trade secret” or “trade secrecy” and 0 otherwise, following Glaeser (2018).

Note: We winsorize all continuous variables at the 1st and 99th percentiles by year with the exception of ρ.

Figure 1: GitHub Adoption Over Time

This figure shows the proportion of firms with an account on GitHub relative to the total percentage of public firms on Compustat from 2008 through 2021. The solid line includes firms from all industries, while the dashed line includes only firms in non-tech industries.

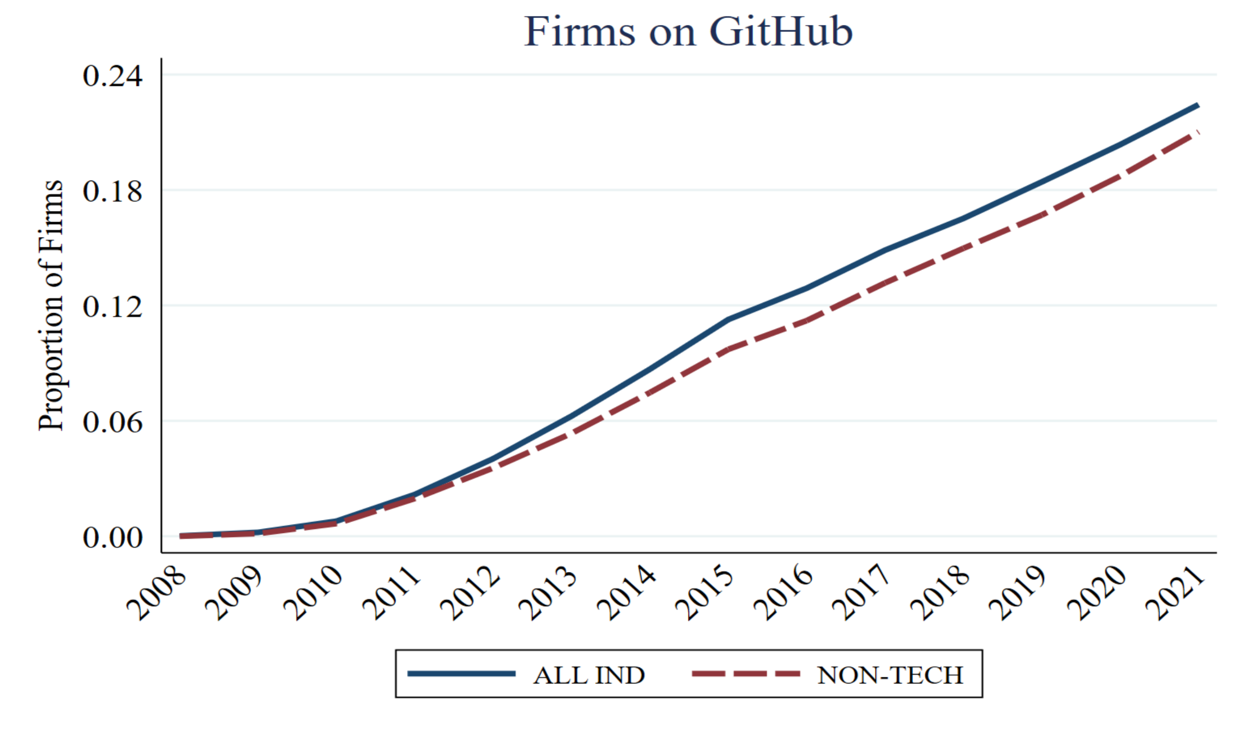


Table 1: GitHub Adoption by Industry and Descriptive Statistics

This table presents the composition of GitHub-adopting firms by industry classification (Fama-French 12). Panel A presents the full sample of firms with a GitHub account during the years 2008-2021 along with their GitHub repository activity and new software patents. Panel B presents information for the sample of firms with a GitHub account during the years 2010-2017. Panel C provides descriptive statistics at the repository level (averaged at the year level across firms' public repositories) for the number of accepted collaborations and forks. Regarding accepted collaborations, "member" refers to individuals who are part of the organization; a "collaborator" is not part of the organization but has defined read, write, or edit permissions for the repository; a "contributor" has made multiple external contributions on a given repository; and the remaining types of collaborations are uncategorized (i.e., "none"). Forks are the number of developer copies of posted public repositories.

Panel A: 2008-2021

Fama-French Industry (12 industries)	**Firm-Years with GitHub Accounts**	**Total Firm-Years on Compustat**	**GitHub Accounts as Perc. of Total**	**Avg. New Repositories (firm-years with at least one)**	**Average New Software Patents**
Consumer Non-durables	266	2,634	10.10	4.87	0.09
Consumer Durables	195	1,460	13.36	3.79	7.49
Manufacturing	331	5,403	6.13	4.34	1.61
Oil, Gas, and Coal	113	3,159	3.58	14.84	0.23
Chemicals	109	1,436	7.59	5.89	0.11
Business Equipment	3,164	11,364	27.86	19.41	15.23
Telephone	241	1,807	13.34	8.65	5.52
Utilities	87	1,538	5.66	2.73	0.06
Wholesale	462	5,116	9.03	28.42	1.17
Healthcare	510	9,822	5.20	4.79	0.13
Finance	574	15,287	3.75	6.59	0.36
Other	695	9,373	7.43	8.28	0.53
Total Observations	*6,747*	*68,399*	*9.87*	*15.24*	*3.24*

Panel B: 2010-2017

Fama-French Industry (12 industries)	**Firm-Years with GitHub Accounts**	**Total Firm-Years on Compustat**	**GitHub Accounts as Perc. of Total**	**Avg. New Repositories (firm-years with at least one)**	**Average New Software Patents**
Consumer Non-durables	87	952	9.14	4.72	0.09
Consumer Durables	67	579	11.57	3.38	5.93
Manufacturing	101	1,825	5.53	4.83	1.20
Oil, Gas, and Coal	39	945	4.13	7.22	0.45
Chemicals	24	589	4.07	8.45	0.18
Business Equipment	989	3,735	26.48	17.92	29.47
Telephone	72	538	13.38	8.51	10.23
Utilities	0	65	0.00	0.00	0.00
Wholesale	153	1,967	7.78	16.15	2.76
Healthcare	98	2,046	4.79	2.00	0.38
Finance	64	4,003	1.60	5.58	0.54
Other	199	3,058	6.51	11.18	0.87
Total Observations	1,893	20,302	9.32	14.41	6.54

Panel C: Community Participation (2010-2017)

	Mean	Std. Dev	25%	Median	75%
Accepted Collaborations	21.627	76.982	0.000	1.052	11.603
Member	3.607	43.677	0.000	0.000	0.122
Collaborator	2.244	11.058	0.000	0.000	0.125
Contributor	14.213	48.991	0.000	0.500	5.941
None	1.563	6.464	0.000	0.000	0.871
Forks	27.190	112.727	0.202	2.400	12.000

Table 2: Summary Statistics

This table presents summary statistics for our main sample. We define all variables in Appendix C. The sample spans from 2010 through 2017 and includes 20,302 firm-year observations. All continuous variables are winsorized at the 1st and 99th percentiles by year.

Variable	Mean	Std. Dev.	25%	Median	75%
AdjNI [+1, +5]	0.570	0.910	0.098	0.478	0.923
NewGitHub Count	0.392	2.589	0	0	0
HighCollab (unlogged)	0.127	1.038	0	0	0
LowCollab (unlogged)	0.250	1.618	0	0	0
HighAdoptFork (unlogged)	0.146	1.138	0	0	0
LowAdoptFork (unlogged)	0.216	1.486	0	0	0
SoftwarePatents	3.072	16.887	0	0	0
NonSoftwarePatents	14.005	64.995	0	0	1
RD	0.036	0.082	0.000	0.000	0.031
TradeSecret	1.332	3.105	0	0	1
MA	0.021	0.056	0.000	0.000	0.008
CAPEX	0.040	0.052	0.006	0.023	0.051
SGA	0.172	0.187	0.031	0.113	0.238
ADV	0.010	0.028	0.000	0.000	0.005
AdjNI	0.083	0.146	0.015	0.086	0.155
ChAdjNI	0.011	0.102	-0.009	0.006	0.033
BM	0.622	0.535	0.283	0.510	0.822
MVE	7,457.769	20,677.533	242.583	1,061.515	4,316.813
AnalystFoll	8.711	8.541	2.000	6.000	13.000
InstOwn	0.598	0.329	0.318	0.691	0.886
Age	20.071	16.600	8.140	16.530	26.058
Lev	0.246	0.254	0.039	0.181	0.373
HHI	0.182	0.124	0.115	0.143	0.196
MGMTFore	0.310	0.463	0	0	1

Table 3: Determinants

This table reports the results of regression analyses that model the determinants of GitHub activity. We cluster standard errors by firm. ***, **, * represent significance at the 1%, 5%, and 10% two-tailed levels, respectively. We define all variables in Appendix C.

Dependent Variable =	*NewGitHub*	
	(1)	(2)
Other Innovative Activity:		
lnSoftwarePatents	0.135***	0.098***
	(9.480)	(7.108)
lnNonSoftwarePatents	-0.013**	0.013**
	(-2.052)	(2.072)
RD	0.345***	0.419***
	(3.600)	(4.063)
TradeSecret	0.004*	0.004**
	(1.922)	(1.976)
MA	0.226***	0.111
	(3.200)	(1.604)
Investment Strategy and Profitability:		
CAPEX	-0.143**	0.011
	(-1.993)	(0.119)
SGA	0.102***	0.058
	(3.123)	(1.548)
ADV	0.637**	0.597**
	(2.217)	(1.989)
AdjNI	0.123***	0.101***
	(3.536)	(2.858)
ChAdjNI	-0.091***	-0.048*
	(-3.483)	(-1.942)
Lev	-0.021	-0.030
	(-1.256)	(-1.591)
BM	-0.008	0.010
	(-1.172)	(1.573)
Information Environment:		
lnMVE	0.021***	0.017***
	(5.449)	(4.253)
lnAnalystFoll	0.015**	0.018**
	(2.040)	(2.520)
InstOwn	0.007	-0.017
	(0.426)	(-1.043)
lnAge	-0.036***	-0.028***
	(-5.543)	(-4.746)
HHI	0.105***	0.111***
	(5.213)	(5.243)
MGMTFore	-0.010	0.005
	(-0.799)	(0.432)
Industry FE	No	Yes
Year FE	No	Yes
Observations	20,302	20,302
Adj. R-squared	0.113	0.183

Table 4: Main Results

This table reports the results of regression analyses of firms' adjusted net income aggregated over the next five years on GitHub activity. We cluster standard errors by firm. ***, **, * represent significance at the 1%, 5%, and 10% two-tailed levels, respectively. We define all variables in Appendix C.

Dependent Variable =		*AdjNI [+1, +5]*	
	(1)	(2)	(3)
NewGitHub	0.338***	0.107***	0.091***
	(11.464)	(5.210)	(4.241)
lnSoftwarePatents		0.004	-0.012
		(0.305)	(-0.819)
lnNonSoftwarePatents		0.037***	0.035***
		(4.605)	(3.692)
RD		0.269	0.110
		(1.162)	(0.431)
TradeSecret		-0.002	-0.002
		(-0.408)	(-0.569)
MA		0.390***	0.304***
		(3.656)	(2.801)
CAPEX		0.138	0.410**
		(0.955)	(2.260)
SGA		0.259***	0.279***
		(3.810)	(3.207)
ADV		3.010***	3.625***
		(6.775)	(7.603)
AdjNI		3.965***	3.869***
		(29.846)	(27.960)
ChAdjNI		-1.001***	-0.988***
		(-10.225)	(-10.003)
Lev		-0.017	0.002
		(-0.439)	(0.033)
BM		-0.146***	-0.141***
		(-9.734)	(-9.129)
lnMVE		0.003	0.013*
		(0.528)	(1.950)
lnAnalystFoll		0.014	0.011
		(1.435)	(1.124)
InstOwn		0.125***	0.115***
		(4.539)	(3.960)
lnAge		-0.022**	-0.016
		(-2.193)	(-1.496)
HHI		0.214***	0.208***
		(4.327)	(4.156)
MGMTFore		0.026	0.024
		(1.620)	(1.486)
Industry FE	No	No	Yes
Year FE	No	No	Yes
Observations	20,302	20,302	20,302
Adj. R-squared	0.026	0.508	0.515

Table 5: Channels of Value Creation

This table reports the results of regression analyses of firms' adjusted net income aggregated over the next five years for each of the cross-sectional tests. Panel A reports the results for high vs. low crowdsource engagement. Panel B reports the results for high vs. low repository adoption. Panel C reports the results for managerial learning. We cluster standard errors by firm. ***, **, * represent significance at the 1%, 5%, and 10% two-tailed levels, respectively, while the significance of the F-test of the difference between *HighCollab* and *LowCollab* is one-tailed. We define all variables in Appendix C.

Panel A: High vs. Low Crowdsource Engagement

Dependent Variable =		*AdjNI [+1, +5]*	
	(1)	(2)	(3)
HighCollab	0.153***		0.120***
	(4.285)		(2.782)
LowCollab		0.097***	0.035
		(3.885)	(1.197)
F-Statistic			
HighCollab – *LowCollab* > 0			0.085*
			(1.778)
Controls	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Observations	20,302	20,302	20,302
Adj. R-squared	0.515	0.514	0.515

Panel B: High vs. Low Adoption

Dependent Variable =		*AdjNI [+1, +5]*	
	(1)	(2)	(3)
HighAdopt	0.157***		0.160***
	(4.628)		(3.746)
LowAdopt		0.077***	-0.005
		(3.195)	(-0.157)
F-Statistic			
HighAdopt – *LowAdopt* > 0			0.165***
			(6.346)
Controls	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Observations	20,302	20,302	20,302
Adj. R-squared	0.515	0.514	0.515

Panel C: Managerial Learning

Dependent Variable =	*AdjNI [+1, +5]*		
	(1)	(2)	(3)
*NewGitHub * RD*	1.380***	1.407***	1.176***
	(5.964)	(5.920)	(2.979)
NewGitHub	-0.008	-0.025	0.106
	(-0.391)	(-1.228)	(0.589)
RD	0.107	-0.096	-0.111
	(0.455)	(-0.372)	(-0.425)
Controls	Yes	Yes	Yes
Industry FE	No	Yes	Yes
Year FE	No	Yes	Yes
NewGitHub * Controls	No	No	Yes
NewGitHub * FE	No	No	Yes
Observations	20,302	20,302	20,302
Adj. R-squared	0.512	0.518	0.521

Table 6: Costs of GitHub Adoption

This table reports the results of regression analyses that examine potential costs of GitHub adoption by regressing the incidence of future cyber security breaches on GitHub adoption. We cluster standard errors by firm. ***, **, * represent significance at the 1%, 5%, and 10% two-tailed levels, respectively. We define all variables in Appendix C.

Dependent Variable =	*Cyber Breach [+1]*	*Cyber Breach [+1, +3]*	*Cyber Breach [+1, +5]*
	(1)	(2)	(3)
NewGitHub	0.005*	0.007*	0.008**
	(1.708)	(1.915)	(2.076)
CyberSecurity_10K	0.007***	0.009***	0.012***
	(3.141)	(3.546)	(4.367)
Controls	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Observations	20,302	20,302	20,302
Adj. R-squared	0.013	0.015	0.015

Table 7: Short Window Returns Following GitHub Repository Postings

This table presents the mean abnormal returns following GitHub repository postings. Panel A presents the results for the sample period 2010-2021. Panel B presents the results for the sample period 2010-2017. In each panel, the first row presents the mean return across all repository postings. The third (fifth) row presents the mean return across high- (low-) adoption repositories. All returns are characteristic adjusted using the matching size, book-to-market, and momentum quintile return (Daniel et al., 1997). The returns window starts the day the repository is created and continues for two days (i.e., three days total). t-statistics are reported in parentheses below each mean return value. ***, **, * represent significance at the 1%, 5%, and 10% two-tailed levels, respectively.

Panel A: Years 2010-2021

Mean Abnormal Returns	N	3-Day (%)
All Repositories	53,820	0.059***
(t-statistic)		(3.43)
High-Adoption Repositories	27,858	0.114***
(t-statistic)		(5.67)
Low-Adoption Repositories	25,962	0.000
(t-statistic)		(0.00)

Panel B: Years 2010-2017

Mean Abnormal Returns	N	3-Day (%)
All Repositories	20,490	0.079***
(t-statistic)		(3.09)
High-Adoption Repositories	10,801	0.117***
(t-statistic)		(3.69)
Low-Adoption Repositories	9,689	0.037
(t-statistic)		(0.91)

Table 8: Robustness

This table presents several robustness analyses. Panel A reports the regression results on various subsamples. Column 1 examines a sample where industry-years (SIC four-digit) have observable software activity. Column 2 examines a sample of only GitHub adopting firms during 2010 through 2017. Panel B presents results of a modified control function estimation (Klein and Vella 2010). In this panel, 90% confidence intervals based on 1,000 bootstrapped samples are in brackets for both *NewGitHub* and *ρ*. Panel C reports alternative ways of measuring GitHub activity. Panel D reports the use of alternative dependent variables. We cluster standard errors by firm in Panels A, C, and D. ***, **, * represent significance at the 1%, 5%, and 10% two-tailed levels, respectively. We define all variables in Appendix C.

Panel A: Alternative Control Samples

Dependent Variable =	*AdjNI [+1, +5]*	
	(1)	(2)
	Indicator of Software Development Activity	Within GitHub Firms
NewGitHub	0.086*** (4.038)	0.047** (2.103)
Controls	Yes	Yes
Industry FE	Yes	Yes
Year FE	Yes	Yes
Observations	16,626	3,426
Adj. R-squared	0.512	0.582

Panel B: Modified Control Function

Dependent Variable =	*AdjNI [+1, +5]*	
	(1)	(2)
NewGitHub	0.094*** [0.037, 0.150]	0.152*** [0.074, 0.230]
ρ	-0.002 [-0.046, 0.041]	-0.144*** [-0.214, -0.074]
First Stage Determinants	Yes	Yes
Controls	Yes	Yes
Industry FE	Yes	No
Firm FE	No	Yes
Year FE	Yes	Yes
Observations	20,302	19,887
Adj. R-squared	0.515	0.824

Panel C: Alternative Independent Variables

Dependent Variable =	*AdjNI [+1, +5]*			
	(1)	(2)	(3)	(4)
ihs_NewGitHub	0.072***			
	(4.222)			
NewGitHub Count		0.015***		
		(4.142)		
NewGitHub Ind			0.135***	
			(3.333)	
ActiveGitHub				0.076***
				(4.447)
Controls	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	20,302	20,302	20,302	20,302
Adj. R-squared	0.515	0.515	0.514	0.515

Panel D: Alternative Dependent Variables

Dependent Variable =	*Sales [+1, +5]*	*AdjNI [+1, +3]*	*AdjNI [+1]*	*Op Earn [+1, +5]*
	(1)	(2)	(3)	(4)
NewGitHub	0.255***	0.033***	0.006***	0.225***
	(3.334)	(3.725)	(3.017)	(4.360)
Controls	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
CC (2011) Controls	No	No	No	Yes
Observations	20,302	20,302	20,302	19,018
Adj. R-squared	0.647	0.577	0.614	0.206