Quid Pro Code: Peer Effects and Productivity in Open Source Software

March 2, 2023
(latest version)

Abstract

We empirically examine the extent to which peer effects influence the private provision of public goods. In the case of public information goods, peer contribution may facilitate or otherwise incentivize further contribution from others, effectively subsidizing private provision. Using the setting of Open Source Software (OSS) contribution, we first utilize a reduced form approach to derive causal estimates of net peer effects in public goods contribution by exploiting a peers-of-peers identification strategy. We next develop a structural model of peer-influenced public good provision that both (1) separates extensive and intensive margin contribution decisions and (2) decomposes contribution into marginal private benefits and costs. We apply these methodologies using a sample of peer contribution histories for 2,287 OSS projects hosted on the GitHub collaboration platform. Both reduced form and structural approaches suggest peer effects are much stronger along the extensive margin than the intensive margin. Contemporaneous intensive margin effects, while heterogenous across time and projects, are small and centered around zero, suggesting that strategic complementarity and substitution in peer contribution likely offset one another. Our counterfactual analysis suggests (extensive margin) peer effects account for nearly 56% of cumulative aggregate contribution for our sample, which translates to a value-added of 1–1.5 million software developer labor hours. These results support the notion that OSS is largely developed by disproportionate efforts from smaller groups of dedicated core maintainers, who integrate incremental contributions from the wider community, and casts doubt on the promise for peer effects alone to deliver sustained maintenance labor to individual projects.

Contents

List of Figures

List of Tables

1 Introduction

Open Source Software (OSS) projects are public information goods produced through incremental efforts of individual contributors.[1][2] Interested parties can freely download software code for their own use and can also propose contributions to the original maintainer of the project[3]. The very existence of OSS rebukes conventional wisdom on privately produced public goods[4] and various explanations have been offered to rationalize their provision, from signalling (Lerner and Tirole, 2002), (impure) altruism (Andreoni, 1990), need satiation (Athey and Ellison, 2014), and institutional structures imposed by self-organizing local communities (Ostrom, 1990; Benkler, 2002). In this study we examine an alternative channel through which widespread contribution to public OSS projects may be achieved: peer effects. Peer behavior can potentially affect the net returns to public good contribution through various channels, improving returns and ameliorating contribution costs. Can peer influence drive heterogeneity in preferences and contribution costs, effectively subsidizing the private provision of public goods?

Consider the quandary faced by maintainers of OSS projects.[5] Sindre Sorhus is a superstar OSS contributor. As of December 2021, he works on OSS full-time and is the author and primary maintainer of over 1,000 OSS projects (Sorhus, 2021). As a prolific maintainer, Sorhus interacts with the wider community of OSS contributors and has personally reviewed tens of thousands of proposed contributions to his projects. Sorhus once reflected that "$\sim$ 80% of contributors doesn't [*sic*] know how to resolve a merge conflict, almost no one writes a good pull request titles, $\sim$ 30% don't run

[1] Our use of the term "open source" requires some definition. In a general sense, OSS is a computer technology for which the underlying source code is made publicly available under a license permitting use, modification, and subsequent redistribution of derived products (Open Source Initiative, 2007). While there are many variations on the specifics of this definition, the most important feature of software projects considered in this study is that (1) they are distributed under some permissive OSS license (GitHub, Inc., 2022) and (2) they are *collaborative* projects that allow for modifications to be submitted from a contributor base wider than the original developer.

[2] Throughout this chapter, we will use the terms "contributor", "developer", "individual", and "agent" interchangeably in reference to the population of study.

[3] For example, a user may wish to propose a new feature or fix a software fault (i.e., a "bug").

[4] Since contribution is costly, agents choose their contribution levels both with respect to private benefits of contribution and the level of the OSS public good delivered by the efforts of their peers. If the net benefit of contribution is negative, an individual may simply opt to free-ride on the efforts of others, leading to misallocation of contribution away from an efficient equilibrium.

[5] In this chapter, we will at times classify agents in the OSS public goods setting according to their level of participation in what is known as the "contributor funnel" (McQuaid, 2018). Users of an OSS project may utilize a software product but do not contribute to it. A subset of users are contributors and allocate some contribution effort to developing the project. A subset of contributors are maintainers, typically agents responsible for a large share of project contribution and may also have decision-making power over what proposed contributions are integrated into the project. We will also sometimes refer to these agents as developers.

tests locally before submitting a [pull request]", and "$\sim$ 40% don't include docs/tests" (Sorhus, 2019). In essence, Sorhus's concern centers around the lack of quality project contributions from his peers. Software development in general is a complex, ever-changing process and many potential contributors simply may lack the skills to contribute effectively. As opposed to shouldering the entire burden of OSS project development[6], to what extent can the contributions efforts of skilled contributors like Sorhus actually improve the productivity of their peers?[7]

A key difference between OSS projects and other public goods is that production of OSS generates both a community of contributors and a set of auxiliary information goods around the project that can potentially reduce subsequent costs of contribution. For example, OSS project maintainers provide assistance and guidance to new contributors by responding to inquiries via mailing lists, message boards, or real-time chat channels.[8] Moreover, OSS communities typically archive the history of such project-related interactions between contributors, creating a publicly accessible knowledge base for project development.[9] OSS projects typically feature documentation[10] that gives a broad overview of the project.[11], provides detailed information on how the software operates at a technical level, and suggest how to properly propose new contributions.[12] Popular OSS projects can also generate a significant amount of buzz outside the contribution platform itself, from community-authored articles demonstrating usage to external forums[13] where users can request help for various programming and software tasks. The combination of these features form the basis for peer effects on contributor productivity. Contribution activity itself can generate a form of "digital capital"[14] for subsequent OSS production, working to both lower the initial fixed

[6]While the use of OSS is itself non-rival, the contribution bandwidth of project maintainers is not (Brown, 2018).

[7]In other words, how do maintainers induce project users down the "contributor funnel" into becoming productive, recurring contributors?

[8]Users who receive feedback on their contribution from project maintainers are far more likely to return to contribute in the future (Sholler et al., 2019).

[9]Similarly, OSS projects are overwhelmingly managed using a *version control system*, making the entire projects incremental development history public record.

[10]Note that documentation is generated by developer labor and a contribution to the project itself.

[11]Examples of high-level documentation include project `README` files bundled with the project source code, "wiki" pages, and long-form vignettes on project usage. For an example of best practices on how these are actually integrated into an OSS project, see Sections 8, 10, and 11 of Wickham (2015).

[12]For example, a project maintainer may include a "contribution template" so that novice contributors avoid common pitfalls for new project contributions. Referring back to the example of Sindre Sorhus, this improves the quality of the proposed change and reduces the "back-and-forth" between maintainer and contributor.

[13]A relevant example is the programming-focused question and answer website Stack Overflow which has been described as a sister community to OSS collaboration platforms such as GitHub (Eghbal, 2020).

[14]Or more accurately, human capital that is recorded or codified as a public information good and then used as an input in the production of additional public goods.

cost of contribution for potential contributors and to make current contributors more productive. Hence, in contrast with many conventional public goods settings, there is scope for individual and peer contribution to become strategic complements.

Salient examples of OSS begin to illustrate the scale at which developers have contributed labor towards the production of complex public information goods. As each OSS developer's "contribution bandwidth" is both scarce and costly, the significance of peer effects that drive contributor labor can be measured naturally in terms of the opportunity cost of a developer's time: what is the equivalent private market labor expenditure to finance the development of large OSS projects? Consider the case of the Linux Kernel. Regarded as the largest collaborative OSS project in history, the Linux Kernel was first released in 1991 by Linus Torvalds and has become the most widely used operating system basis for web servers, mobile devices, and high performance computing infrastructure. As of September 2021, the Linux Kernel has amassed over 31.3 million single lines of code from 23,927 distinct contributors over the past three decades. Using standard methods from software engineering cost estimation, it would take nearly 70 million person-hours to rewrite the entire kernel from scratch, which would cost over $1.05 billion today.[15][16] While estimates for the use-valuation of OSS is an important ongoing area of research (Greenstein and Nagle, 2014; Nagle, 2019), in this study we seek to characterize the extent to which peer effects can mitigate production costs of OSS public goods.

We seek to empirically assess peer effects on public good production using the context of OSS. Our methodology is organized into two phases. In the first phase, we build intuition on the magnitude of net peer effects in OSS contribution using a reduced form approach. To address concerns over endogeneity, we develop an identification strategy to determine to what extent individual effort levels are influenced by the contribution levels of their peers. Specifically, we instrument the likely endogenous contribution effort of an agent's peers in a given project with the effort levels of the agent's "peers-of-peers" defined by common contribution in outside projects.[17] The instrument operates by changing the relative incentives for peers to contribute to a given project by varying the incentives in external projects. This approach allows us to determine whether individual and peer

[15] Estimated (conservatively) using the COCOMO model of software development cost estimation developed by Boehm (1981) and the software utility `scc` (Source: https://github.com/boyter/scc).

[16] The median annual salary for software developers in the United States for 2020 was $110,140 ($52.95 per hour) (U.S. Bureau of Labor Statistics, 2021).

[17] Details for this identification strategy are given in Section **??** and Figure 1.

contribution are strategic complements or substitutes on net, conditional on the set of developers that contribute at all. In the second phase, we develop a structural model of OSS contribution to pin down the microeconomic foundations for contributor behavior. We seek to place emphasis on disentangling contribution decisions along the extensive versus intensive margin and integrate peer influence into both decisions. To this end, we embed a micro-founded model of private public good provision (Bergstrom et al., 1986) into the selection model of (Heckman, 1979). The structural approach facilitates the recovery of individual productivity parameters, allowing us to characterize the welfare of particular contribution profiles and conduct counterfactual analysis. Our main counterfactuals of interest estimates the value of aggregate contribution added by peer effects.

We apply this methodological framework in an empirical analysis, focusing on the context of Open Source Software contribution. We use individual-level contribution data for a random sample of 2,287 highly collaborative OSS projects hosted on the GitHub collaboration platform. The remainder of this chapter is organized as follows. We first provide additional background on OSS development in Section **??**. Next, we survey segments of related literature in Section **??**. We introduce the empirical setting in Section **??**, describing OSS contribution activity on the GitHub platform and giving an overview of data included in the empirical sample. We then develop a reduced form strategy to estimate peer effects in Section **??**. With high-level insight on net peers effects in hand, we next develop a structural model of public good contribution with extensive and intensive margin peer effects in Section **??**. We outline an estimation strategy, present estimation results, and conduct counterfactual analysis to measure the value of contribution generated by distinct peer effects channels. Finally, we summarize and interpret our findings in Section **??** and discuss promising directions subsequent research.

2 Background

OSS projects are typically organized around *software code repositories*, publicly accessible websites that host the project's source code and provide collaboration functionality[18]. Users can view and

[18]For example, Figure 2 depicts a snapshot of the web user interface for the `bootstrap` project's GitHub repository, a popular JavaScript framework for web development: https://github.com/twbs/bootstrap. The history of all user contributions to the `pandas` codebase can be viewed here: https://github.com/twbs/bootstrap/commits/main. The repository contains a README with the source code, a document that contains links to detailed documentation, installation and usage notes, and guidance for prospective contributors.

download the source code of OSS projects for their own use. They can also contribute to the OSS project's codebase. A typical contribution pattern works as follows: (1) a user downloads a copy of the source code, (2) makes a series of incremental changes to the codebase, and (3) submits a request to the owner of the original OSS repository to integrate their changes. Due to the open nature of the code and the permissiveness of OSS licenses in general, there is little to prevent a user from simply copying the codebase of an existing project into a new OSS good[19]. However, users can distribute contribution costs and share knowledge by working collaboratively with peers. It is therefore reasonable to assume there exist strong motivations for distributed users to rally around and contribute to particular OSS projects instead of splintering off into isolated endeavors, creating "digital communities" around OSS projects characterized by social norms and stocks of project-specific information capital.

Peer effects have long been discussed as a driving force behind the "success" of particular OSS projects. An early discussion on net effect of peer influence on open software contribution began with the conjecture by Brooks Jr (1995), who observed that the addition of developers to a software project slows down the pace of development. In a response to the so-called "Brook's Law", Raymond (1999) countered this postulate with the example of OSS collaboration and "Linus's Law" that roughly states that the likelihood that faults in a software's codebase will be identified and fixed rises with the number of users and contributors working with it. Raymond (1999) argues that the proliferation of highly collaborative, decentralized OSS projects is itself a rebuke of Brook's Law.

As the production of OSS can clearly be subject to peer influence, a core impetus for this study is to disentangle the various channels through which peer effects operate and estimate the empirical implications for these effects on equilibrium contribution. In theory, peer effects can have both negative and positive impacts the level of privately provided OSS. What anecdotal evidence do we have for either (1) free-riding or (2) productivity externalities in OSS development? With the rise of OSS use, a common concern amongst OSS project maintainers is over-subscription of their projects: users who flood communication channels with support requests without contributing the fix themselves (Eghbal, 2020). A related concern is that many OSS projects originating from small

[19]This process is known as "forking" in the OSS community. Forks of original projects can also become active contribution communities in their own right. This typically happens when there is a sufficient number of contributors interested in pursing a different direction of development.

groups of contributors are widely used as part of the "digital infrastructure" (Eghbal, 2016) that underpins modern information and communication technologies. Consider the case of OpenSSL, an encryption library that by some estimates is used by two-thirds of public facing web servers to secure private information (The OpenSSL Project Authors, 2021). In 2011 a bug, now known as Heartbleed, was introduced into the OpenSSL codebase and was not discovered until 2014[20], exposing a vast swath of internet communications that were previous thought to be secure. The estimated cost to simply limit the extent of this vulnerability was estimated to be over $500 million USD (Kerner, 2014) and does not consider the cost of any secure data lost through the exploit. The OpenSSL team at the time "never had more than three to four core developers" overseeing more than a half a million lines of code on an annual donation budget of $2,000 USD (Oberhaus, 2019). Whether it was the sheer size and complexity the OpenSSL codebase or the preferences of the maintenance team deterred potential contributors, the fact that an OSS project serving as a critical component of internet infrastructure did not receive more attention from the wider community of users who rely on it ought to be cause for concern for OSS sustainability.

While free-riding on OSS contribution is likely prevalent and perhaps inescapable when considering a project's user-base in the broadest sense[21], it may also be the case that increased participation in OSS distributes the joint cost contribution and improves individual productivity. How can the development of OSS itself either make subsequent contribution less costly or induce the marginal free-rider to contribute? Recommended practices in software engineering encourage developers to include documentation, testing frameworks, and use automated processes whenever possible (Fogel, 2005). Documentation explains the functionality and inner workings of software code in plain language, making it easier for both users and potential contributors to work with the software. Testing frameworks ensure the code functions as intended and are essential for a large collaborative OSS project. Continuous Integration (CI), a form of automation in the integration and testing of changes to software projects, facilitates a greater volume of contribution and has been shown to allow software projects to release[22] more frequently (Hilton et al., 2016). Investments in these features lower the cost burden of maintenance and lower the barriers to entry for

[20]Consequently, some have pointed to the Heartbleed exploit as a repudiation of Linus' Law (Meneely et al., 2014).

[21]Modern software projects, both proprietary and open source, typically borrow 70 to 90% of their functionality OSS components (Nagle et al., 2022).

[22]In software development, a "release" is a particular version of the project distributed to users. In an appeal to Linus' Law, OSS proponents such as Raymond (1999) and Fogel (2005) encourage frequent releases.

new contributors. Moreover, active contributors in OSS communities often provide "non-code" contribution services to the project, answering user inquiries, reviewing and integrating proposed changes, establishing design principles and community guidelines, and other functions peripheral to contributing code. It's natural to imagine that all else equal, a potential contributor would prefer allocating their contribution bandwidth to an OSS project with sociotechnical infrastructure that makes it easier to work with.

The collaborative and decentralized nature of OSS development suggests a setting rife with intricate peer effects. The wider population of OSS users may lack the skills or resources needed to contribute to OSS codebases and may simply free-ride on the contributions of more prolific developers. At the same time, OSS contribution itself generates an abundance of features that reduce the cost of and further incentivize wider OSS participation. We use this study as an opportunity to develop a microeconomic framework decomposing these forces and to empirically estimate their implications.

3 Related Literature

We review a subset of academic literature that can be divided into several distinct strands: (1) motivations for OSS contribution, (2) the private provision of public goods, and (3) peer effects.

3.1 Why Contribute to OSS?

Although initially puzzling, the existence and proliferation of OSS goods has been studied through an economic lens for over two decades (Lerner and Tirole, 2002). A common interest in early research on the economics of OSS focuses on the incentives for participation in public good production by both individuals and profit-maximizing firms. Different hypotheses have been offered to explain OSS provision and contribution behavior:

- Individual private benefits: intrinsic motivation (Lakhani and Wolf, 2003), need satiation (Bessen, 2006; Athey and Ellison, 2014), signalling and status (Glazer and Konrad, 1996; Lerner and Tirole, 2002; Roberts et al., 2006), "warm glow" (Andreoni, 1990), option value of modular codebases Baldwin and Clark (2006), permissive licensing (Fershtman and Gandal, 2004; Lerner and Tirole, 2005; Fershtman and Gandal, 2007).

- Social effects: pure altruism (Bonaccorsi and Rossi Lamastra, 2003), social norms and reciprocal altruism (Raymond, 1999; Bergquist and Ljungberg, 2001; Benkler, 2002), project productivity (Fershtman and Gandal, 2011)
- Strategic motivations for firms: innovation, market power (Bonaccorsi et al., 2006), labor search[23], cost reduction (Andersen-Gott et al., 2012)

Some closely related work examines contribution to OSS and open source content in general empirically. Fershtman and Gandal (2004) find that permissive software licenses induce greater levels of contribution. Hahn et al. (2008) find that OSS developers are more likely to join projects with past collaborators. Fershtman and Gandal (2011) demonstrate an empirical relationship between the success of an OSS project, measured in downloads, and the extent to which its contributors work in other common projects, suggesting the existence of both direct and indirect project knowledge spillovers. In contrast, the present study uses microdata to measure peer effects on contribution at the individual level. Several authors have used the context of Wikipedia to study peer effects within collaborative production of open content. Exploiting blockages of Chinese language Wikipedia for mainland China, Zhang and Zhu (2011) find that pro-social peer effects are increasing in the number of peers: individuals contribute more to Wikipedia when they have more peers. Slivko (2014) use an indirect peers strategy to find modest evidence for positive, intensive margin peer effects amongst frequent contributors.

3.2 Private Public Good Provision

Seminal work seeks to rationalize private provision of public goods. While the canonical public goods model of Samuelson (1954) suggests strong incentives to free-ride on the contributions of others, heterogeneity in both preferences and the marginal cost of provision can explain positive levels of private provision in many contexts (Tiebout, 1956; Stiglitz, 1981, 1982; Bergstrom et al., 1986; Cornes and Sandler, 1985; Andreoni, 1990; Fischbacher and Gächter, 2006; Kotchen, 2009; Jacobsen et al., 2017). In the case of OSS, online collaboration dramatically reduces transaction costs inherent to the production of other types of public goods (Coase, 1937; Nitzan and Romano, 1990). Social norms develop around projects in order to efficiently manage the needs of the community

[23]See https://github.com/t9tio/open-source-jobs for a list of job listings for private firms with primary products centered around GitHub OSS repositories.

and the time constraints faced by contributors (Holländer, 1990; Ostrom, 1990). Moreover, agents are subject to contribution externalities and can confer productivity benefits on peers, which in turn confer additional benefits to the original agent (Elliott and Golub, 2019). In this sense, agents "pass through" benefits of increased contribution and can be compensated for these investments.

Several authors have focused on public good provision specifically within the context of OSS. Johnson (2002) analyzes a model of OSS public good contribution. As expected, the assumption of the fixed costs of contribution preclude the efficiency of the decentralized equilibrium. Baldwin and Clark (2006) find that highly "modular" codebases provide contributors with option value and ultimately attract more contribution.

3.3 Peer Effects

Productivity Spillovers

Particularly of concern to our reduced form analysis, we link this work to an expansive body of literature concerning peer effects and their estimation. Experimental evidence suggest peer effects in public goods settings can be driven by punishment (Fehr and Gächter, 2000), cooperation (Falk and Ichino, 2006), and can ultimately increase voluntary contribution to public projects (Archambault et al., 2016). Several empirical studies find evidence of labor productivity "spillovers" when high ability peers are introduced (Mas and Moretti, 2009; Lindquist et al., 2015). There is mixed evidence for peer effect heterogeneity across individuals (Arcidiacono and Nicholson, 2005; Cornelissen et al., 2017), suggesting the context and estimation strategy matter. A related literature investigates the importance of group sizes on treatment and peer effects (Angrist and Lavy, 1999; Krueger, 2003).

Identification

Identification of peer effects in non-experimental settings is of great concern to this literature. Manski (1993) posits a "reflection problem" which Bramoullé et al. (2009) suggest can be solved by using instruments generated by the network structure itself: the behavior of indirectly linked agents can generate quasi-random variation needed to address endogenity concerns with estimating peer effects in observational data. It should be noted that the identification strategy of Bramoullé et al. (2009) relies purely on characteristics of the network structure between agents and without

qualification, can be devoid of microeconomic foundations or even lack appealing quasi-random variation for causal identification (Angrist, 2014). Other authors have used alternative strategies, such as true random assignment of peers (Sacerdote, 2001; Guryan et al., 2009; Carrell et al., 2011), exploiting quasi-experimental designs (Dahl et al., 2014), overlapping peer groups (De Giorgi et al., 2010), directly modelling endogenous peer networks (Goldsmith-Pinkham and Imbens, 2013), the use of panel data (Patnam, 2011), and explicit structural approaches (Ciliberto et al., 2016). Our study will draw several techniques from this literature to develop an identification strategy for peer effects, including social connections, changes in peer groups, and individual fixed effects to develop a unique, micro-founded "peers-of-peers" identification strategy in Section **??**. To the best of our knowledge, the closest use of the peers-of-peers identification strategy in public good contribution is Slivko (2014), who uses the number and average contribution level of indirect peers to instrument for peer contribution.

4 Data

We use observational data to measure individual contribution levels over time for a sample of OSS projects. We draw this sample from projects hosted on GitHub, the world's largest collaborative software development platform.[24] For each project, we observe agent-level contribution efforts in continuous time[25], measured in "commits", or atomistic modifications to the codebases of OSS projects.[26] For the purposes of this study, we will define an agent's peer group in a given project as the set of other developers contributing code to that project. Individual and peer contribution levels across projects and time will form the core of the reduced form and structural analysis. Additional details on the data used in this paper can be found in Section **??** of the appendix.

We begin by describing the dataset in broad strokes. Since the universe of OSS repositories on the GitHub platform is incredibly vast[27], we restrict our empirical sample to a randomly selected

[24] Launched on April 2008, GitHub has become the world's largest source code host and *de facto* collaboration platform for OSS projects

[25] Each commit to a project is recorded with a timestamp (e.g. 2009-10-31 01:48:52).

[26] Note that a commit can encompass changes to any number of lines across any number of files. A natural concern may be that variation in the size of individual commits makes it difficult to compare as equivalent units of contribution effort. For example, a single commit might be a simple typo correction requiring little effort or a complicated "bug" fix that took many hours to address. Some have argued for simpler measures to estimate labor commitment to software development, such as the number of days a developer makes at least one contribution to a project in a given time period (Sherwood, 2015).

[27] As of January 2020, GitHub has over 40 million users and hosts more than 190 million software repositories

subset of popular and highly collaborative projects. Specifically, we take a 10% random sample of GitHub projects with 15 or more distinct contributors and 100 or more "stars"[28] as of June 2019. This results in an empirical sample containing 2,287 projects and 107,921 distinct contributors observed from the launch of GitHub in April 2008 through June 2019.[29] We aggregate individual contribution to a monthly frequency and therefore the unit of analysis is individual-project-time.[30] The most commonly represented programming languages for these projects are JavaScript (31%), Python (11%), and Java (9%). Of the contributors represented in the sample, 3.7% are members of the projects they contribute to and only 0.57% are project owners.[31] The average project in the empirical sample is 5 years old and is the product of 2,490 cumulative commits made by 56 distinct contributors. The average individual in the sample contributes 13 commits to a particular project a month and 53 commits across all projects over the sample period. It is critical to note the (right) skewness of contribution, both between and within projects: the median project has 829 cumulative commits made by 29 distinct contributors while the median agent makes only 3 commits to a single project each month. Furthermore, the share of individual contribution within projects is bimodal (see Figure 4). Roughly 45% of contributions in our sample are made by agents who represent 5% or less of total project contribution for that month. On the other end of the spectrum, about 8% of observations in the sample represent individual contributions that account for over 95% of total project commits for that month. In simpler terms, the most common contribution pattern within projects involves many individuals contributing a small share[32] relative

(GitHub, Inc., 2020).

[28]On the GitHub platform, users can mark interesting projects by "starring" them, which subscribes the user to a newsfeed covering project development. For the purposes of this study, we use project stars as a proxy for user interest or quality of the project. Stars also distinguish highly collaborative, "engineered" software projects from small, single-user projects (e.g., abandoned forks or repositories containing personal files like notes or school projects) (Munaiah et al., 2017).

[29]Since GitHub is simply the web platform hosting the project, some projects in the sample have contributions made either prior to the existence of GitHub or it's arrival on the GitHub platform. Projects are managed using version control systems (VCS) that record a complete history of changes in the project since its inception. GitHub's namesake comes from the VCS tool used by the projects it hosts: `git`. Figure 3 overlays histograms for (1) the earliest recorded commit in each project and (2) the date the project was created or moved to the GitHub platform.

[30]In other words, each observation is the level of contribution by an individual to a particular project for the given month.

[31]For the projects in this sample, "ownership" does not imply property rights over the software code itself. Project "ownership" and membership on the GitHub platform simply means the user has certain administrative privileges within the repository, most important of which is the ability to merge proposed contributions of outsiders into the main project codebase. It should be noted that many projects may feature core contributors with a considerable amount of influence on project design decisions who are not officially project owners or members in the GitHub system.

[32]This is known as "drive-by" or "casual" contribution (Fogel, 2005; Eghbal, 2020).

to a dominant core contributor in each period. The sample provides evidence that even though both aggregate contribution and the number of distinct contributors have grown over time (see Figure 5 and Figure 6), average individual contribution levels have remained quite stable (see Figure 7). Consistent with anecdotal evidence from the OSS literature (Eghbal, 2020) and theory on contributor behavior (Athey and Ellison, 2014), these characteristics suggest the growth of an OSS projects is a combination of (1) small number of dominant core contributors and (2) the aggregate effect of small contributions from a wider population of software developers.

With a general understanding of the GitHub contribution sample used in this chapter, we now direct attention towards measures of peer and individual contribution germane to both reduced form and structural analysis. We present key descriptive statistics for this empirical sample of contributions measures in Table 1. The average agent contributes 13 commits to a project each month and has an average of 17 peers contributing 188 commits in aggregate. As noted before, individual contribution is highly right-skewed. The median agent contributes just 3 commits per month and has 7 peers who contribute 59 total commits. Approximately 6.8% of observations in the empirical sample involve a sole contributor with no peers in that time period. Since the mean and median individual contribution levels coincide with project-specific contribution, these data suggest that most agents contribute to a single project in a month.[33] The average agent's mean cumulative contribution to a particular project is 256 commits (median 23), a pattern that naturally is similar in peers. Together, insights from the empirical sample suggest agents form affinities with a particular project and continue to contribute to it over time.[34]

Finally, we collect two additional measures most relevant for our structural approach described in Section **??**. First, for each project and time period, we measure the number of "stars" associated with the project. This is a rough proxy for an OSS repository's popularity and is used to measure the level of public good quality. Similar to contribution levels, project quality is highly skewed: the

[33]It should be noted that the apparent lack of contributors contributing to multiple projects may simply be an artifact of sample construction. We simply take a random sample of projects and observe contribution activity of agents within those particular projects. Therefore, individuals in the sample may be contributing to other OSS projects not recorded in the present sample. We at least partially account for this deficiency when constructing our instrument for peer contribution in Section 5.2, a measure that sums contribution levels of "peers-of-peers" across all projects recorded in the GHTorrent sample of Gousios (2013), some of which are not contained in our empirical sample.

[34]This may be explained by many alternative mechanisms, including individual need (Bergstrom et al., 1986; Lerner and Tirole, 2002; Lakhani and Wolf, 2003), the discipline of social norms (Ostrom, 1990), or an accumulated expertise within a project.

mean (median) project has 910 (161) stars in a given month. Second, we observe individual time allocated on the platform. Specifically, for each individual and time period, we measure how many days they spend contributing to any project on the GitHub platform (Sherwood, 2015). We use this measure of time allocation to proxy for numéraire good consumption, which in turn facilitates the estimation of time and project-varying productivity shocks for each agent. In a given month, the average (median) agent makes commits over 3.88 (2) days to projects in the sample. Compared with the skewness in contribution levels, this descriptive suggests a stark difference between the extensive and intensive margin contribution decisions.

5 Reduced Form

Before developing a structural model, we build intuition for net peer effects in public goods contribution using a simple reduced form framework. We seek to understand how an individual agent's individual contribution level is influenced by the contribution level of her peers. This section is organized as follows. We first outline a baseline econometric specification to assess peer effects in public good contribution. Next, in an effort to address endogeneity concerns and give a causal interpretation to the peer effect estimates, we propose an instrumental variable for peer contribution, define its measurement, and discuss various possible threats to identification. The final subsection discusses the empirical results.

5.1 Peer Effects on Individual Contribution

Consider a setting in which individuals $i \in \mathcal{N}$ contribute to OSS projects $p \in \mathcal{P}$ in each period $t \in \mathcal{T}$. The outcome of interest, $a_{ipt} \geq 0$, is the contribution level for agent i to project p at time t. The aggregate contribution level of agent i's peers to project p at time t, denoted by $a_{\text{-}ipt} \equiv \sum_{j \neq i} a_{jpt}$, is the regressor of interest. We present a baseline specification[35] for contribution peer effects in Equation (1):

$$a_{ipt} = \delta a_{\text{-}ipt} + \boldsymbol{\beta}' \boldsymbol{X}_{ipt} + \epsilon_{ipt}. \tag{1}$$

[35] Alternative specifications similar to Equation 1 are presented in Section 5.4, serving to both provide robustness checks and to consider different characterizations of peer influence.

Here the vector $\boldsymbol{X}_{ipt}$ is a set of observable exogenous factors driving agent i's level of contribution to project p at time t. The term ϵ_{ipt} represents unobservable factors driving contribution.

The coefficient of interest in Equation (1) is δ, which captures the (average) effect of aggregate peer contribution on the level of individual contribution.[36] We will refer to the coefficient δ as the reduced form peer effect in contribution. This term is sometimes referred to in the literature as the "endogenous effect" (Manski, 1993) or "social multiplier" (Glaeser et al., 2003). Our empirical analysis seeks to test the null hypothesis of no peer effects ($\delta = 0$) against an alternative that there exists some relationship between contribution levels of peers ($\delta \neq 0$). If there is evidence of peer effects, we are also interested in the net effect of the opposing externalities. The core premise of this study is that peer influence is the net effect of two distinct externalities in contribution. In the canonical public goods model, individual and peer contribution to public goods are strategic (gross) substitutes and therefore voluntary provision is vulnerable to free-riding. If incentives to free-ride dominate, we should expect $\delta < 0$[37] in equilibrium. On the other hand, if an increased level of peer contributions also leads to an increase in agent i's contribution *ceteris paribus*, it is likely the case that some other peer effect (e.g., externalities in productivity and contribution costs, pro-social behavior) dominates incentives to free-ride. This would imply $\delta > 0$.

Other observable factors that influence agent contribution are captured in a vector $\boldsymbol{X}_{ipt}$ and may potentially vary across agents, OSS projects, and time. Examples of these influences may include individual and peer contribution history[38], observable quality or popularity of the OSS project, the size of the contribution peer group, technical characteristics of the projects[39], and other agent characteristics.[40][41][42] In terms of the specification in Equation (1), we can also include

[36] In other words, the effect on individual contribution when peer contributions increase by 1 commit, on average and *ceteris paribus*.

[37] Note that this assumes that $\boldsymbol{\beta} \neq \mathbf{0}$ in the true model for individual contribution. If the population model in Equation (1) is such that $\boldsymbol{\beta} = \mathbf{0}$ (i.e., a model without covariates or an intercept), then $\delta \in [0, 1]$ by construction. As $\delta \to 0$, a single contributor dominates and all others free-ride. As $\delta \to 1$, contribution is uniform across peers.

[38] Such as an agent's cumulative contribution to a project at time t or their contribution in previous periods. Cumulative and temporal lags of contribution can capture an agent's accumulated experience or affinity with a particular project.

[39] Such as project age and programming language used

[40] In the context of GitHub data available, agent characteristics may include whether the agent is the owner or member of the project or if they identify with a particular employer.

[41] Agents can voluntarily include the name of their employer in their GitHub profile and can make contributions with a company email address.

[42] On the GitHub platform, an agent can be added as a member to a project, potentially giving them more discretion over what proposed changes by the wider community are integrated. It also is plausibly a signal of an agent's affinity with a particular project.

a battery of individual, project, or time fixed effects[43] in $\boldsymbol{X}_{ipt}$.

5.2 Identification

The specification in Equation (1) describes a model in which peer groups are defined as the set of agents contributing to a particular OSS project at a given point in time: individual contribution is a function of contemporaneous, aggregate peer contribution. The specification is a simplified linear-in-sums[44] formulation similar to reduced form models studied widely in the peer effects literature (Manski, 1993; Bramoullé et al., 2009; Goldsmith-Pinkham and Imbens, 2013).

Point identification of the parameter δ in Equation (1) is demonstrated by Lee (2007) by exploiting "leave out" sums and variation in peer group sizes, overcoming the well known non-identification result of Manski (1993). In our setting, we point to the descriptive statistics in Table 1 as evidence that contributors are likely to have different groups of contributing peers in each period for any particular project. Since peer groups in the current empirical setting are naturally quite dynamic, we argue that point identification is established. Therefore, we wish to go one step further and establish causal identification for the parameter δ in Equation (1). Under what conditions can we interpret an estimate of δ as the local average treatment effect (LATE) of peer contribution on the level of individual contribution? The immediate challenge is that since individual and peer contribution are both observed choice variables, a naive estimate of δ likely suffers from endogeneity bias (Angrist, 2014; Lewbel, 2019). An experimental ideal to causally identify the net peer effects parameter δ would involve first randomly assigning agents to projects and then allowing them to decide contribution levels, ensuring random peer groups in which choice of contribution levels ought to be uncorrelated unobservables, or $\text{Cov}(a_{\text{-}ipt}, \epsilon_{ipt}) = 0$. In reality, agents select into and choose contribution levels on the basis of potentially unobservable influences, such as personal need, technical ability, and their endowment of time to work on OSS projects. For example, high ability agents might select into and make above-average contributions to common projects, generating positive bias in the ordinary least squares estimate of δ in Equation (1) (i.e., $\text{Cov}(a_{\text{-}ipt}, \epsilon_{ipt}) > 0$). On the other hand, low ability agents may also select into projects with highly skilled developers and make minimal contributions. Taken to the extreme, agents who free-ride

[43] Inclusion of individual level fixed effects $\boldsymbol{X}_{ipt}$ accounts for an agent's intrinsic proclivity to contribute to OSS goods, independent of other factors (Andreoni, 1990).

[44] Technically, the specification is linear in "leave out" sums by definition of $a_{\text{-}ipt}$.

completely do not contribute at all and therefore do not appear in the sample whatsoever. Since we cannot completely account for all free-riders in OSS[45], we must acknowledge any interpretation of the estimated effects in this analysis is to be conditional on the population individuals who contribute at all.

In the absence of purely random assignment of contributors to projects, we address the concern over endogeneity in peer contribution $a_{\text{-}ipt}$ by use of an instrumental variables strategy. We seek a valid instrument for peer contribution that, conditional on other observable and exogenous factors $\boldsymbol{X}_{ipt}$, (1) exerts some influence on the contribution levels of Agent i's peers $j \neq i$ and (2) only influences Agent i's contribution to project p through its effect on peer contribution $a_{\text{-}ipt}$.[46] In other words, in lieu of random assignment to projects, we must opt for an instrument that generates quasi-random variation in peer contribution levels, conditional on the set of agents that contribute at all to a given project. Furthermore, we combine this instrumental variables approach with a battery of both control variables that plausibly explain OSS contribution and fixed effects at the individual, project, and time period to account for common but unobservable shocks across each unit.[47]

Contribution by peers-of-peers

Consider an agent i who contributes to OSS project p. If for some reason i's peers suddenly find contribution to other projects relatively more (less) attractive, they may allocate efforts away from (towards) project p through a channel with no direct influence over Agent i's contribution to p. This strategy is facilitated by the project-mediated "social network" of individual developers in which connections are defined by the projects they commonly contribute to.[48] Agent i has peers $j \neq i$ in project p, who in turn also have peers $k \neq i, j$ in other projects $q \neq p$ they also contribute to. Hence, we argue we can use the contribution network structure itself in a "peers-of-peers" identification strategy in the spirit of Bramoullé et al. (2009) to recover the effect of peer effort

[45] Given its wide-reaching prevalence, it's difficult to imagine there exist consumers of information technology who have *not* used OSS at some point.

[46] In the language of instrumental variable estimation, an instrument that satisfies both the (1) relevance and (2) exclusion conditions.

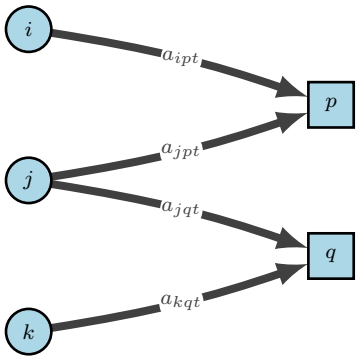
[47] In the context of our notation, across each i, p, and t.

[48] In a similar effort, Fershtman and Gandal (2011) use a bipartite graph to model connections between OSS projects and contributors.

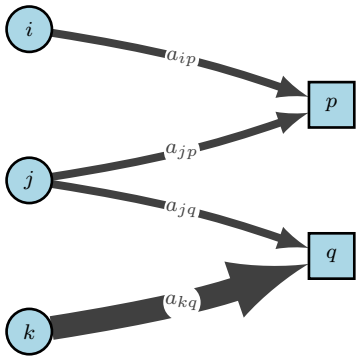
levels on equilibrium contribution levels.[49] An important departure is that while the strategy of Bramoullé et al. (2009) is designed to exploit general characteristics of the peer social network, the identification used in this study is based on microeconomic principles of substitution. We sketch out the identification strategy graphically in Figure 1. To guide the graphical intuition, consider the following hypothetical scenario. Suppose there are three contributors $\mathcal{N} = \{i, j, k\}$ and two OSS projects $\mathcal{P} = \{p, q\}$. Assume that at the beginning of period t, contribution profiles are $a_{ipt}, a_{jpt}, a_{jqt}, a_{kqt} > 0$. Hence, Agents i and j contribute positive amounts to Project p while Agents j and k contribute positive amounts to Project q. Agent i's direct peer is Agent j and indirect or "peer-of-peer" is Agent k. In this sense, Agents i and k are connected only indirectly through the

Figure 1: Identification Strategy ("Peers-of-peers" Contribution). Agents $\{i, j, k\}$ contribute to Projects $\{p, q\}$. Assume i and j contribute to p while j and k contribute to q.

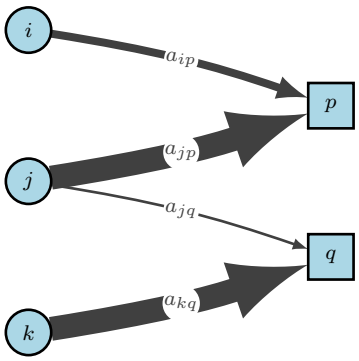
(a) Initial Setting



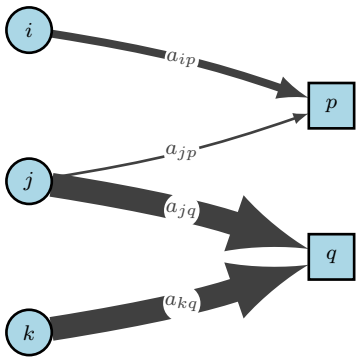
(b) Suppose Agent k increases contribution to Project q



(c) Case 1: Agent j substitutes *towards* Project p (contribution is a strategic substitute)



(d) Case 2: Agent j substitutes *away from* Project p (contribution is a strategic complement)



[49] As noted when surveying related literature, Slivko (2014) uses a similar "peers-of-peers" identification strategy by using a network of Wikipedia editors mediated by articles commonly contributed to.

contribution patterns of Agent j.[50] This initial setting is represented in Panel (a) of Figure 1. Next, suppose Agent k increases their contribution to Project q (e.g., Panel (b) of Figure 1). If changes in Agent k's contribution to project q influence the time-constrained contribution behavior of Agent j, then Agent j may have incentives to change her contribution levels to Project p. The case in which Agent j finds her contribution to Project p a strategic complement with Agent k's is depicted in Panel (c) of Figure 1. An example in the OSS setting may occur when Agent k contributes a fix for an issue in Project q that was consuming Agent j's contribution bandwidth. Conversely, the case in which Agent j finds her contribution to Project p a strategic complement with Agent k's is depicted in Panel (d) of Figure 1. This may arise if Agent k contributes an attractive fix or feature to Project q that encourages additional contribution from Agent j. In either case, the contribution pattern of Agent i's indirect peer Agent k influences Agent i only through changes in the behavior of Agent j.

In summary, we propose the use of aggregate contribution of peers-of-peers effort to instrument for peer effort. The instrument operates by inducing substitution of contribution effort across projects, generating quasi-random variation in aggregate peer effort from the perspective of individual developers. In the following two subsections, we define the measurement of the peers-of-peers instrument and provide a set of assumptions for its validity.

Instrument Measurement

Denote the "peers-of-peers" instrument for peer contributions $a_{\text{-}ipt}$ as z_{ipt}. Roughly speaking, we choose to define z_{ipt} as the aggregate contribution of peers of i's peers in project p at time $t-1$.[51] Formally we measure z_{ipt} as t:

$$z_{ipt} = \sum_{j \neq i} \sum_{q \neq p} \sum_{k \neq i,j} 1\{a_{jq,t-1} > 0\} 1\{a_{jk,t-1}\} a_{kq,t-1}. \tag{2}$$

Hence, z_{ipt} represents "aggregate contribution by i's peers-of-peers defined by project p in month $t-1$".[52] To avoid concerns of reverse causality, we use peers-of-peers contribution in the previous

[50] That is to say, Agents i and k are not *directly* connected through the contribution networks. Any influences Agent k's contribution has on that of Agent i operate *only* through changes in Agent j's behavior.

[51] We consider the contribution of peers-of-peers in the previous period to mitigate concerns over reverse causality.

[52] Since the sample is constructed by measuring all contribution around a set of core collaborative projects, it is important to note that agents in the empirical sample also contribute to outside projects. Therefore, projects $q \neq p$

month $t-1$ to instrument for peer contribution in month t. Since Equation (1) postulates that agents respond to aggregate as opposed to average peer contribution, we construct the peers-of-peers instrument similarly.

Threats to Identification

The validity and strength of the peers-of-peers instrument z_{ipt} instrument for peer contribution $a_{\text{-}ipt}$ rests on several assumptions.

Assumption 5.1. *No isolated contributors.*

Most obviously, contributors need to have peers in order to assess the influence of peer effects. Moreover, their peers must also have peers. While by construction of the sample each project has at minimum 15 distinct contributors over its lifespan, we acknowledge that the empirical sample includes a small share of observations in which only a single agent makes a contribution in that time period.[53] We should reasonably expect such observations to both weaken the relationship between the instrument z_{ipt} and peer effort a_{ipt} and introduce downward bias to estimates of the peer effect coefficient δ.

Assumption 5.2. *For each agent i, there exists a set of projects i will never contribute to, independent of the cost of contribution.*

Agents cannot be peers with everyone. For the exclusion restriction to hold, it is necessary that peers-of-peers contribution influences individual contribution only through peer contribution. Hence, we need a sufficient level of contribution behavior where agents are connected indirectly through peers.[54] Consider a setting in which all agents contribute to all projects. In this setting, agent i's contribution level is directly influenced by other agents since the "peers of i's peers" are really just i's peers.

Assumption 5.3. *Conditional on observable influences, agents substitute contribution effort between projects.*

and contribution levels $\{a_{jqt}, a_{kqt}\}$ may not be present in the sample.

[53] As noted in Section **??**, these observations comprise about 6.8% of the empirical sample.

[54] In other words, the social network of contribution cannot be a complete graph and a sufficient number of "intransitive triads" exist in the contribution network (Bramoullé et al., 2009; De Giorgi et al., 2010; Graham, 2015).

In other words, the peers-of-peers effort z_{ipt} is conditionally correlated with aggregate peer contribution $a_{\text{-}ipt}$ for the relevance condition to hold: $\text{Cov}(a_{\text{-ipt}}, z_{ipt} \mid \boldsymbol{X}) \neq 0$. This is our most critical assumption. For the instrument to be relevant, there needs to exist some degree of influence of peers-of-peers contribution on peer contribution in aggregate. The immediate concern with using "peers-of-peers" contribution to instrument for peer effort in Equation (1) is that the objective of reduced form analysis is to test the null hypothesis of $\delta = 0$: no influence of peer contribution on individual contribution within projects. However, it is important to note that the peers-of-peers instrument operates through substitution with peer contribution *between* projects while the null hypothesis for Equation (1) only accounts for substitution with peer effort *within* projects. There is no reason *ex ante* that one substitution pattern precludes the other. Additionally, we argue that the peers-of-peers contribution levels are relevant conditional on other exogenous or predetermined factors drive peer contribution, such as cumulative contribution in a project (i.e. $\text{Cov}(a_{\text{-}ipt}, z_{ipt} \mid \boldsymbol{X}_{ipt}) \neq 0$).[55] Combining these arguments, we assert that if these assumptions hold then the peers-of-peers instrument z_{ipt} drives some degree of meaningful variation in peer contribution $a_{\text{-}ipt}$ that is quasi-random and therefore exogenous from the perspective of the individual i.

5.3 Results

We present baseline estimates for the peer effects parameter δ of the reduced form model from Equation (1) in Table 2. Columns (1) through (3) of Table 2 present ordinary least squares (OLS) estimates while Columns (4) through (6) present instrumental variables estimates using two-stage least squares (IV 2SLS). We use peers-of-peers contribution z_{ipt} to instrument for peer effort $a_{\text{-}ipt}$. Columns (1) and (4) estimate a specification of Equation (1) with only an intercept and the endogenous regressor $a_{\text{-}ipt}$. Columns (2) and (5) add covariate controls[56] and Columns (3) and (6) add covariate controls alongside project and year-month fixed effects.

The estimates in Table 2 suggest little evidence of peer effects in contribution on average for the full sample. The OLS estimates in Column (3) are not statistically different from zero after

[55] Furthermore, we note there is some degree of mechanical correlation between $a_{\text{-}ipt}$ and z_{ipt}: a greater number of peers to individual i is likely to subsequently generate a greater number of peers-of-peers.

[56] Control variables include three temporal lags of individual and peer commits to the project, cumulative individual and peers project commits, project quality measured in GitHub stars, quadratic terms for project age, and dummy variables indicating if the individual is a project owner, project member, or if they are affiliated with a firm.

accounting for project fixed effects and observables. The same is true for the corresponding 2SLS estimates. These specifications explain roughly 18% of the variation in individual contribution for the full sample. We note the F statistic from the first stage of the 2SLS estimate in Column (6) of Table 2 is 64.37.[57] Hence, given the model in Equation (1) and the sample at hand, we cannot reject the null hypothesis of no peer influence on individual contribution on average ($\delta = 0$) once we account for project fixed effects and covariate controls.

5.4 Detailed Analysis and Robustness

Estimates for the population average δ in Equation (1) mask considerable heterogeneity in peer influence on individual level contribution. To explore heterogeneity and provide additional robustness for the reduced form peer effect estimates, we estimate a series of alternative specifications and present the results in Appendix **??**. Most notably, we find evidence that although the number of contributors has grown over time, contemporaneous peer effects have diminished over time (see Figure 9). It's reasonable to suspect that peer effects are stronger in the earlier days of GitHub as most projects in the sample were still in their infancy. Peer groups were smaller and there were simply fewer developers active on the platform. We also find considerable heterogeneity in peer effects at the project level (see Figure 8). We interpret this result as heterogeneity in net complementarity of contribution effort that likely varies across projects.[58] Moving beyond contemporaneous peer effects, we find that peer effects are stronger when regressing individual contribution 3 months after on peer contribution 3 months preceding a given period t (see Table 5). It is likely the case that peer influence takes some time to operate and individuals are induced to contribute on the basis of relatively recent development activity, not necessarily occurring in the same month. This result is important as it suggests that intensive margin peer effects are likely stronger after relaxing our rather restrictive assumption of contemporaneous influence.[59] Finally, we consider the effect of con-

[57] Since the model is just-identified, we report the heteroskedasticity-robust F statistic proposed by Olea and Pflueger (2013) and recommended by Andrews et al. (2019).

[58] Moreover, our data and econometric specification treat each individual commit as equivalent contribution. In reality, some contributions might be more important than others. Consider the difference between a typo fix and the introduction of a new feature set. Since the specific lines of code changed by each commit can be observed, a worthwhile continuation of this work ought to examine the interaction between specific types of contribution and peer influence. This approach can give context to the heterogeneity observed in peer effect estimates and guide recommendations for OSS sustainability policy.

[59] We return to this assumption when interpreting our structural estimates of intensive margin peer effects in Section 6.5.

tribution by project "insiders" on the level of contribution by project "outsiders" (see Figure 10), and find evidence that increased contribution from project insiders "crowds out" contribution from outsiders.

Our results together suggest that while contemporaneous intensive margin peer effects in contribution are limited on average for the entire sample, (1) there exists significant heterogeneity in peer effects across time and projects, (2) contemporaneous peer effects may too narrowly restrict the scope of peer influence, and (3) free-ridership is likely prevalent if dominant core contributors and project insiders carry out the bulk of OSS development. We discuss these results in more detail alongside the findings of the structural analysis in Section **??**.

6 Structural Model

While the reduced form analysis begins to reveal patterns of peer influence in OSS contribution, a more refined approach is needed to operationalize the various channels through which peer contribution can influence equilibrium behavior. Importantly, the reduced form specification in Equation (1) conflates peer influence along both the extensive and intensive margin into a single parameter. Therefore, estimates of contribution peer effects δ are conditional on agents who contribute positive amounts and does not separately account for *why* agents decide to contribute to particular projects. A structural approach allows us to rigorously define micro-founded channels for both equilibrium contribution decisions and peer influence.

There are several key features of our structural model. First, we seek to separately identify marginal private benefits and costs of contribution for each agent. Second, we can characterize each agent's equilibrium contribution decision along both the extensive (i.e., whether to contribute) and intensive (i.e., how much to contribute) margins. Third, we can integrate peer influence into both of these features. Peer effects can potentially influence contribution benefits and productivity as well as intensive and extensive margin contribution in equilibrium. Finally, a fully specified structural approach permits counterfactual analysis. This will allow us to place a value-added estimate for both intensive and extensive margin peer effects in terms of changes to equilibrium contribution. Specifically, we can compare contribution from the observed equilibrium to a counterfactual under

which peer effects are absent.[60]

The remainder of this section is organized as follows. First, we set up the model of OSS contribution and introduce its various elements in Section 6.1. Our approach combines a model of private provision of public goods (Bergstrom et al., 1986) into a selection model (Heckman, 1979). Second, we define an equilibrium in Section 6.2. Third, in Section 6.3 we specify how peer effects enter into the structural framework. Fourth, we detail our estimation strategy in Section 6.4. Fifth, we describe the structural estimates in Section 6.5. Finally, we conduct a counterfactual analysis to estimate "value-added" by peer effects in Section 6.6.

6.1 Setup

Individual agents (i.e., OSS developers) are indexed $i \in \mathcal{N} = \{1, \ldots, N\}$. In each period $t \in \mathcal{T} = \{1, \ldots, T\}$, agents choose contribution levels $a_{ipt} \geq 0$ across a set of OSS projects indexed $p \in \mathcal{P} = \{1, \ldots, P\}$ to maximize incremental contribution utility in each period. To summarize what follows, Table 8 collects notation for the structural model.

Project Quality

Projects are indexed by their quality y_{pt} at time t. We assume project quality y_{pt} is a simple linear function y of cumulative contribution through t, $a_{pt} \equiv \{a_{ips}\}_{s \leq t}^{i \in \mathcal{N}}$, and parameters b_{pt}:

$$y_{pt} = y(a_{pt}, b_{pt}) = b_{pt} \sum_{i \in \mathcal{N}} \sum_{s \leq t} a_{ips}. \tag{3}$$

Note that this specification implies that the parameter b_{pt} represents the marginal product of contribution labor in terms of the quality of project p at time t.[61]

Preferences

Agent preferences are styled after Bergstrom et al. (1986)'s model of private public good provision. We extend this framework to include multiple public goods and time periods. In each period t and

[60] Or, as if software developers contributed to projects in isolation from one another. We operationalize this by setting intensive or extensive margin peer effects equal to zero.

[61] While the project quality specification in Equation 3 may give rise to concerns over "over-fitting" parameter estimates to the data, we choose this specification purposefully to capture the reality that the marginal product of contribution labor is arguably higher when the project is in early development stages.

for each project p, agents derive utility over (1) direct contribution benefits, (2) project quality, and (3) a numéraire consumption good x_{it} (e.g., time). Specifically,

$$u_{it} = u(a_{it}, y_t, x_{it}) = \sum_{p \in \mathcal{P}} \left(v_{ipt} a_{ipt} - \frac{1}{2}(a_{ipt})^2 + y_{pt} \right) + x_{it}, \tag{4}$$

where $a_{it} \equiv \{a_{ipt}\}_{p \in \mathcal{P}}$ and $y_t \equiv \{y_{pt}\}_{p \in \mathcal{P}}$ respectively collect agent i's contributions and project quality across for all $p \in \mathcal{P}$ at time t. Following Bergstrom et al. (1986), we assume linear preferences: contribution, public good quality, and private good consumption are perfect substitutes.[62] This simplifies the utility maximization problem into independent choices of optimal contribution between projects, subject only to a budget constraint. Agent preferences are shaped by private contribution benefit shocks $v_{ipt} \in \mathbb{R}$, which partially determine the optimal level of contribution in equilibrium. It's critical to note that a realization of v_{ipt} may be such that the agent decides not to contribute to project p at all. Individual project-specific benefit shocks are similar to Athey and Ellison (2014)'s "arrival of needs" model of OSS dynamics at a macro-level.

Contribution Constraint

We assume that agent contribution a_{ipt} and consumption of the private good x_{it} are constrained by (1) productivity shocks[63] $c_{ipt} > 0$ and (2) endowments ω_{it}:

$$x_{it} + \sum_{p \in \mathcal{P}} c_{ipt} a_{ipt} \leq \omega_{it}. \tag{5}$$

In our empirical application, ω_{it} is the agent's endowment of time in period t (i.e. 1 month) and the private good x_{it} is the amount of time spent *not* contributing.[64] As in the reduced form analysis, we measure a_{ipt} as the number of commits agent i makes to project p at time t. This implies that the (inverse) productivity parameters c_{ipt} measure the time cost incurred by agent i making a_{ipt} commits to project p.[65] If $c_{ipt} > c_{jpt}$, agent j is *more* productive contributing to project p at time

[62] More specifically, preferences are quasilinear in x_{it} and therefore increasing an agent's endowment of the numéraire good does not influence demand for contribution.

[63] As specified in the contribution constraint in Equation (5), c_{ipt} technically represents agent i's "cost" of contribution to project p at time t. The inverse of c_{ipt} is therefore a measure of contribution productivity. Throughout the structural analysis, we will refer to c_{ipt} as both "productivity" and "cost" interchangeably.

[64] In other words, the number of days in month t in which agent i authored no commits.

[65] When an agent's endowment ω_{it} is measured as the number of days in period t and x_{it} is measured as the number of days in the period i was not active on the GitHub platform, the shock c_{ipt} can be interpreted as the number of

t. Finally, we naturally normalize $\omega_{it} = 1$ for all i and t given its interpretation. Since $a_{ipt} \geq 0$, this will in turn imply $0 \leq x_{it} < 1$ and $0 < c_{ipt} < 1$.[66]

Selection Mechanism

Obviously, agents can elect to contribute nothing to certain projects. We therefore introduce a selection mechanism in the spirit of Heckman (1979). We assume that projects feature fixed costs of contribution, modelled as a latent productivity threshold z_p.[67][68] Agent i will contribute $a^\star_{ipt} > 0$ to project p at time t if their private project-specific ability z_{ipt} exceeds z_p. Furthermore, we assume that z_{ipt} is a linear function of observables $\boldsymbol{W}_{ipt}$, $z_{ipt} = \boldsymbol{\gamma}'\boldsymbol{W}_{ipt} + \epsilon^z_{ipt}$ where $\epsilon^z_{ipt} \sim \mathcal{N}(0, 1)$. Therefore, the probability that i contributes to project p in period t is

$$\Pr(a^\star_{ipt} > 0) = \Pr(z_{ipt} \geq z_p) = \Phi(\boldsymbol{\gamma}'\boldsymbol{W}_{ipt}), \tag{6}$$

where $\Phi(z)$ is the standard normal cumulative distribution function. In applications, we normalize the contribution threshold $z_p = 0$ for all projects.[69] The vector $\boldsymbol{W}_{ipt}$ contains a set of characteristics that influences i's decision to contribute to project p at time t: the number of peers contributing to project p and both cumulative and lagged contribution for individual i as well as for all agents $j \neq i$ (e.g., historical peer contribution). These factors give important signals to prospective contributors deciding which projects to participate in and will serve as the basis for our extensive margin peer effects discussed in more detail in Section 6.3. For example, an established project featuring many active contributors can provide a useful signal to newcomers uncertain about its quality and maturity, who may be more inclined to contribute under a belief that their efforts will go towards a worthwhile endeavor.

commits i makes to project p per days i was active on GitHub.

[66] In general, we only bound productivity shocks such that $c_{ipt} > 0$. However, in the data the smallest value of positive contribution is $a^\star_{ipt} = 1$. It can therefore be shown that this normalization implies also $c_{ipt} < 1$ for all $a^\star_{ipt} > 0$.

[67] We acknowledge that this selection mechanism could also be interpreted as a latent *benefit* threshold. See the discussion of structural estimates of extensive margin peer effects in Section 6.5.

[68] See Hsieh et al. (2018) and Hsieh et al. (2020) for examples of similar selection mechanisms used in models of public good contribution.

[69] To rationalize this normalization, we detail project-specific estimation of $\boldsymbol{\gamma}$ for each p in Section 6.3.

6.2 Equilibrium

Timing and Information

At the beginning of each period t, each agent first learns their extensive margin shock ϵ^z_{ipt} for each project. Next, the set of agents who meet the productivity threshold $z_{ipt} \geq z_p$ and decide to contribute to project p learn their benefit and productivity shocks (v_{ipt}, c_{ipt}). We assume that all shocks are public information: agents know who will contribute to which project and how much they will contribute. In the following subsections, we characterize both extensive and intensive margin decisions and the resulting equilibrium.

Extensive Margin Decision

Following the selection mechanism described in Equation (6), agent i will contribute $a^\star_{ipt} > 0$ if and only if $z_{ipt} \geq z_p$ upon learning ϵ^z_{ipt}. Otherwise, if an agent does not cover the productivity threshold, they will decide not to contribute to project p at all: $z_{ipt} < z_p \iff a^\star_{ipt} = 0$.

Intensive Margin Decision

Agents with $z_{ipt} \geq z_p$ next determine an optimal, positive contribution level $a^\star_{ipt} > 0$. Taking marginal private benefit and productivity shocks (v_{ipt}, c_{ipt}) as given, each agent i chooses an allocation $(a_{ipt}, y_{pt}, x_{it})$ to maximize incremental utility u_{it}:

$$\begin{aligned} \max_{a_{ipt}>0, y_{pt}, x_{it} \in [0,1)} \quad & \sum_{p \in \mathcal{P}} \left(v_{ipt} a_{ipt} - \frac{1}{2}(a_{ipt})^2 + y_{pt} \right) + x_{it} \\ \text{s.t.} \quad & x_{it} + \sum_{p \in \mathcal{P}} c_{ipt} a_{ipt} \leq 1 \\ & y_{pt} = b_{pt} \sum_j \sum_{s \leq t} a_{jps}. \end{aligned} \tag{7}$$

Under the intensive margin decision characterized by System 7, each agent i explicitly takes into account (1) shocks (v_{ipt}, c_{ipt}) and (2) cumulative contribution to project p. To account for affinities and experience formed in particular projects, we allow an agent's cumulative and lagged contribution history to influence their benefit and productivity shocks in Section 6.3.

To characterize each agent's intensive margin contribution behavior in equilibrium, we observe

that the first order necessary conditions for optimal, non-zero contribution $a^{\star}_{ipt} > 0$ imply

$$a^{\star}_{ipt} = b_{pt} + v_{ipt} - c_{ipt}. \tag{8}$$

In other words, should agent i decide to contribute to project p at time t, her optimal level of contribution equals the sum of the marginal product of labor in terms of public good quality b_{pt}, the marginal private benefit of contribution v_{ipt}, and the marginal private cost of contribution c_{ipt}. All else equal, agents contribute more when either their marginal product of labor or marginal private benefits of contribution are higher and less when the marginal cost of contribution (i.e., inverse productivity) is higher.[70]

Combining the optimal intensive margin choice of contribution in Equation (8) and the extensive margin decision (i.e., selection mechanism) in Equation (6), a given agent i's equilibrium contribution strategy for project p at period t can be summarized as

$$a^{\star}_{ipt} = \begin{cases} b_{pt} + v_{ipt} - c_{ipt} & \text{if} \quad \boldsymbol{\gamma}'\boldsymbol{W}_{ipt} \geq \epsilon^{z}_{ipt} \\ 0 & \text{if} \quad \boldsymbol{\gamma}'\boldsymbol{W}_{ipt} < \epsilon^{z}_{ipt}. \end{cases} \tag{9}$$

6.3 Peer Effects

We allow peers to influence equilibrium contribution decisions along both the extensive and intensive margins for equilibrium contribution behavior described in Equation (9). To disentangle these margins, we will assume separate channels of influence for each mechanism. Historical peer contribution will form the basis for peer effects along the extensive margin. Conditional on the set of agents who contribute a strictly positive level, correlation between the realized benefit and productivity shocks, (v_{ipt}, c_{ipt}), of an individual and her peers will form the basis for peer effects for the intensive margin choice. We formalize these peer effect channels in the following two subsections.

Extensive Margin

To integrate peer influence into the extensive margin contribution decision, we disaggregate influences over agent i's latent ability threshold for project p at period t, $z_{ipt} = \boldsymbol{\gamma}'\boldsymbol{W}_{ipt} + \epsilon^{z}_{ipt}$, into

[70] Equation (8) is a linear form of the optimal public good contribution level of derived by Bergstrom et al. (1986), reflecting that private public good contribution is driven by heterogeneity in both benefit and cost heterogeneity.

characteristics specific to i or project p, $\boldsymbol{\beta}_z'\boldsymbol{X}_{ipt}$ (individual controls), and those related to peers $j \neq i$, $\boldsymbol{\gamma}'\boldsymbol{W}_{ipt}$ (peer influences).[71] Specifically, we include (1) the number of agents contributing to project p in period $t-1$ as well as (2) cumulative and lagged peer contribution to project p in the vector $\boldsymbol{W}_{ipt}$. This is designed to capture the fact that past contribution to OSS projects by peers is a public information good itself and may lay the foundation for subsequent contribution.[72] On the other hand, agents may also choose to free-ride should cumulative project contribution reach a particular level. The vector $\boldsymbol{X}_{ipt}$ contains measures such as individual i's cumulative and lagged contribution, and therefore accounts for i's own accumulated experience with project p. Furthermore, we allow parameter vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_z$ to vary by project and period. In addition to simplifying estimation[73], estimating separate parameters for each project implicitly accounts for project-varying characteristics that may influence selection beyond contribution history.[74] For each project, the selection mechanism in Equation (6) becomes

$$\Pr(a^{\star}_{ipt} > 0) = \Phi(\boldsymbol{\gamma}'\boldsymbol{W}_{ipt} + \boldsymbol{\beta}_z'\boldsymbol{X}_{ipt}). \tag{10}$$

The parameter vector $\boldsymbol{\gamma}$ captures project-specific peer effects along the extensive margin as a function of historical peer contribution activity. If $\boldsymbol{\gamma} > 0$, the likelihood of contribution is increasing in past peer contribution $\boldsymbol{W}_{ipt}$.[75]

Intensive Margin

For each $a^{\star}_{ipt} > 0$, we can separately recover the shocks v_{ipt} and c_{ipt} by using the equilibrium contribution level in Equation (8), the budget constraint in Equation (5), and the project quality function in Equation (3).[76] Therefore, we can develop a framework for assessing contemporaneous peer influence for both individual private benefits and productivity along the intensive margin,

[71] With a slight abuse of notation for simplicity.

[72] The influence of past actions by peers is also considered in a similar fashion by Bollinger and Gillingham (2012), who use cumulative solar panel installations in a neighborhood to predict current period adoptions.

[73] Estimating an analogous model $\Pr(a^{\star}_{ipt} > 0) = \Phi(\boldsymbol{\gamma}'\boldsymbol{W}_{ipt} + \boldsymbol{\beta}_z'\boldsymbol{X}_{ipt})$ with a single coefficient $\boldsymbol{\gamma}$ would entail a single regression with $N \cdot P \cdot T = 107,250 \cdot 2,287 \cdot 134 = 32,867,620,500$ observations at the individual level.

[74] This amounts to including a distinct constant term in each N-length vector $\boldsymbol{X}_{ipt}$ for each p.

[75] Notice that

$$\frac{\partial \Pr(a^{\star}_{ipt} > 0)}{\partial \boldsymbol{W}_{ipt}} = \boldsymbol{\gamma}\phi(\cdot) > 0$$

[76] Estimation is covered in detail in Section 6.4 as well as Section **??** of the appendix.

conditional on the set of agents with strictly positive contribution levels. In the context of our model, this can be measured by the degree to which shocks (v_{ipt}, c_{ipt}) are correlated between peers in a given project and period.

We separate peer effects in contribution productivity, c_{ipt}, from peer effects in private contribution benefits, v_{ipt}, by using distinct peer effect specifications similar in structure to the reduced form peer effects specification in Equation (1). First, we assume that agent productivity is at least partially determined by peer effects:

$$c_{ipt} = \delta_c \bar{c}_{\text{-}ipt} + \boldsymbol{\beta}_c' \boldsymbol{X}_{ipt} + \epsilon_{ipt}^c, \tag{11}$$

where $\bar{c}_{\text{-}ipt} \equiv \frac{1}{n_{pt}-1} \sum_{j \neq i} 1\{a_{ipt}^\star > 0\} c_{jpt}$ and $n_{pt} \equiv \sum_{i \in \mathcal{N}} 1\{a_{ipt}^\star > 0\}$ define the mean of productivity shocks for i's contemporaneous peers in project p.[77] Like the extensive margin specification in Equation (10), $\boldsymbol{X}_{ipt}$ are a vector of observables and fixed effects such as lagged and cumulative contribution. Conditional on covariates, δ_c captures the average correlation in productivity shocks amongst peers for a given project and time period. When $\delta_c < 0$, individual costs of contribution are negatively correlated with peer costs, suggesting positive peer effects in terms of productivity.

Similarly, private benefit shocks (e.g., private "needs" or returns to contribution) are modelled as follows:

$$v_{ipt} = \delta_v \bar{v}_{\text{-}ipt} + \boldsymbol{\beta}_v' \boldsymbol{X}_{ipt} + \epsilon_{ipt}^v. \tag{12}$$

When $\delta_v > 0$, individual private benefits are positively correlated with those of their peers, suggesting pro-social peer effects.

To summarize, extensive margin peer effects are parameterized by $\boldsymbol{\gamma}$. Conditional on the set of agents who do contribute, intensive margin (contemporaneous) peer effects are parameterized by (δ_c, δ_v).[78] The framework for extensive and intensive margin peer effects in this structural approach captures several desirable properties. First, we model each margin independently, allowing us to estimate them separately. The source of extensive margin effects is historical peer contribution

[77] When there is only a single agent contributing to a project, $\bar{c}_{\text{-}ipt} = 0$.

[78] In a more general sense, elements in the vectors $(\boldsymbol{\beta}_v, \boldsymbol{\beta}_c)$ may include terms related to historical (i.e., lagged or cumulative) peer contribution that, similar to the extensive margin parameterization, may also plausibly influence positive contribution levels. With respect to the counterfactual analysis in Section 6.6, we are more broadly interested in parameters related to both types of peer influence, contemporaneous and accumulated, on private benefits and productivity.

and the source for intensive margin effects is contemporaneous correlation with contributing peers. Second, given that we observe each agent's extensive margin decision for every project and period, we can estimate $\boldsymbol{\gamma}$ separately for each p. Motivated in part by the considerable project-level heterogeneity revealed in the reduced form analysis, this parameterization is more flexible than estimating a single parameter and can account for a range of project-varying extensive margin influences.[79] Finally, we use the OSS contributor's time-constrained utility maximization problem to separately recover benefit and productivity shocks. Unpacking net benefits allows us to further isolate the channels of peer influence in intensive margin contribution. In the next section, we turn our attention to estimating parameters of interest.

6.4 Estimation

In this section, we provide a high-level overview of our structural estimation strategy and objectives. A more thorough and detailed treatment is provided in Section **??** of the appendix. Given data $(a_{ipt}, y_{pt}, x_{it})$ for all $i \in \mathcal{N}, p \in \mathcal{P}$, and $t \in \mathcal{T}$, we develop an estimation strategy to recover the following:

1. Marginal product of labor parameters $\boldsymbol{b} = (b_{pt})$ from the project quality function in Equation (3).
2. Private benefit and productivity shocks $\boldsymbol{s} = (v_{ipt}, c_{ipt})$ for all $a^{\star}_{ipt} > 0$ from the equilibrium contribution level in Equation (8).
3. (Extensive margin peer effects) Parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta}_z)$ from Equation (10).
4. (Intensive margin peer effects) Parameters $(\delta_c, \delta_v, \boldsymbol{\beta}_c, \boldsymbol{\beta}_v)$ from Equations (11) and (12).

The parameters of interest are $\boldsymbol{\delta} = (\delta_c, \delta_v)$, which drive intensive margin peer effects, and $\boldsymbol{\gamma}$, which drive extensive margin peer effects. For each project $p \in \mathcal{P}$, our estimation strategy is as follows:

1. Assume disturbances are jointly normally distributed $(\epsilon^z_{ipt}, \epsilon^v_{ipt}, \epsilon^c_{ipt}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, independent and identically distributed between agents and time. Within the variance-covariance matrix $\boldsymbol{\Sigma}$, assume that $\sigma^2_z = 1$.
2. Given data (a_{ipt}, y_{pt}), recover b_{pt} using Equation (3).

[79] Note that given data limitations, we cannot estimate intensive margin peer effects (δ_v, δ_c) separately for each project and period. In many cases, our empirical sample contains only a single contribution for the month to a given project.

3. Given data $(a_{ipt}, y_{pt}, x_{it})$ and b_{pt}, recover shocks (v_{ipt}, c_{ipt}) using Equation (9), Equation (5), Equation (3) by way of generalized method of moments (GMM) estimation[80].
4. Given data $(1\{a_{ipt} > 0\}, \boldsymbol{W}_{ipt}, \boldsymbol{X}_{ipt})$ and shocks (v_{ipt}, c_{ipt}) recover $(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, where $\boldsymbol{\delta} = (\delta_v, \delta_c)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_z, \boldsymbol{\beta}_v, \boldsymbol{\beta}_c)$, via maximum likelihood estimation (MLE) (Zhao et al., 2020).

Parameters $\boldsymbol{\theta} = (\boldsymbol{b}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ allow us to completely characterize the data generating process for the structural model, a necessary prerequisite simulating policy counterfactuals.

6.5 Structural Estimates

Benefit and Productivity Shocks

We present the recovered values for marginal product of labor parameters b_{pt} and shocks (v_{ipt}, c_{ipt}) for all $a^{\star}_{ipt} > 0$ in Figure 11. The first panel of Figure 11 contains distributions of marginal private benefit shocks v_{ipt} grouped by year. Similarly, productivity (inverse marginal cost) shocks are presented in the second panel. Several patterns emerge from the recovered shocks. First, these distributions are relatively stable over time. Second, when considering the entire sample, benefit and productivity shocks are relatively uncorrelated with one another at the individual level ($\text{Corr}(v_{ipt}, c_{ipt}) = -0.081$). There is, however, evidence of a temporal trend in shock correlation over the sample period: Figure 12 reveals that benefits v_{ipt} demonstrate a strong negative correlation with productivity c_{ipt} ($\text{Corr}(v_{ipt}, c_{ipt}) \approx -0.6$ to -0.5) in early periods of GitHub that trend towards 0 nearer the end of the sample period. Recall that $\text{Cov}(v_{ipt}, c_{ipt}) < 0$ implies that greater marginal private benefits are associated with lower private marginal costs of contribution. Together, these data seem to suggest that early stages of GitHub OSS collaboration featured highly productive individuals with greater net benefits of contribution relative to later entrants. In later periods, incentives to become more productive may be weaker given greater peer participation. Corroborating the findings of the reduced form analysis, this structural evidence further supports the notion that the prevalence of free-ridership has likely increased on average as the GitHub platform has grown in size.

The third panel of Figure 11 contains estimates of marginal product of labor parameters b_{pt}

[80] For each i and t, there are $2P$ unknowns: v_{ipt} and c_{ipt} for each $a_{ipt} > 0$. There are P first order conditions from Equation (9), P equations for project quality form Equation (3), and one budget constraint. Overall, this implies $NT(2P+1)$ moment conditions and $2NPT$ unknowns.

from Equation (3). By virtue of functional form assumption for project quality, b_{pt} tend to be largest in the early stages of project development: the initial commits tend to be the most important in determining project quality. Since b_{pt} tends to decline over a project's lifespan, productivity and benefit shocks explain sustained contribution.

Extensive Margin Peer Effects

Figure 13 contains estimates for extensive margin peer effects captured by the parameter $\boldsymbol{\gamma}$ of Equation (10). Two key patterns emerge. First, the likelihood of contribution is increasing in the number of peers who contributed in the previous period while decreasing in lagged and cumulative contribution levels. Second, the coefficients for lagged number of peers are much larger in magnitude compared with lagged and cumulative contribution. Taken together, these estimates underscore an intuitive if not trivial fact: agents are more likely to join projects growing in the number of contributors. To a lesser extent, the likelihood of contribution declines as projects grow larger in terms of the size of the codebase. We can interpret this finding in several ways. On one hand, actively developed projects provide positive peer effects that incentivize contribution from outsiders. On the other, it may simply be the case that increased development activity in the early stages of a project may signal a project's promise or quality to prospective contributors. To rule out this signalling mechanism, we estimate extensive margin peer effects at the project level and control for observable project quality. Moreover, it appears that contribution incentives lessen as a project matures into a stable state[81], as it is likely that less contribution is required.

In Equation (6) of Section 6.1, we model extensive margin selection into projects as a latent productivity threshold. We acknowledge that the largest driver of project participation, the number of peers contributing, can influence both benefits and costs of contribution. Given that z_{ipt} is unobserved and a function of both individual and peer historical contribution, we could just as easily have modelled z_p as a latent benefit threshold for project p. At best, we can only say our structural approach finds evidence that projects with many actively contributing members increase an individual's *net* benefit of contribution and therefore positively impacts extensive margin participation.

[81] This phase of project development is sometimes referred to as "maintenance mode" as opposed to "active development".

Intensive Margin Peer Effects

Project-level estimates of the intensive margin peer effects δ_v and δ_c are summarized in Figure 14. Much like the project-level reduced form estimates displayed in Figure 8, both benefit and productivity peer effects are distributed relatively symmetrically around 0. A relatively strong positive correlation between δ_v and δ_c, $\text{Corr}(\delta_v, \delta_c) = 0.843$, implies greater benefit correlation between peers within projects is also associated with greater marginal cost correlation between peers. Ultimately, this suggests an *inverse* relation between benefit and productivity shocks correlation: the net effect of peer influence along the intensive margin leads developers to contribute more at greater marginal cost. The lack of correlation between v_{ipt} and c_{ipt} at the individual level further supports this finding. We interpret this positive correlation between δ_v and δ_c as evidence that pro-social peer effects dominate productivity peer effects. Consistent with the reduced form analysis, there is no strong evidence that contemporaneous peer effects improve intensive margin productivity across projects on average. In other words, we cannot say that OSS contributors make each other more productive along the intensive margin when considering contemporaneous influence.

Summary

To summarize, structural estimation of benefit and cost shocks along with extensive and intensive margin peer effects seem to corroborate evidence from our reduced form approach and descriptive statistics from the empirical sample. First, extensive margin peer effects are a much more important driver of project growth relative to intensive margin effects. Consistent with the "casual contributor" phenomenon described anecdotally by OSS maintainers, projects with many contributors are more likely to attract incremental contributions from outsiders than they are to attract dedicated maintainers. Second, in terms of the ratio of private contribution benefits to costs, early OSS contributors on the GitHub platform enjoyed greater net benefits of contribution relative to later entrants. Finally, pro-social forces seem to trump peer effects with respect to intensive margin productivity. There is little evidence to suggest peers reduce marginal costs of contribution along the intensive margin.

It is important to note that these results and their subsequent interpretation rest on some assumptions made in our modelling approach. First, as in the reduced form analysis, we place

a restrictive assumption that intensive margin peer influence operates contemporaneously. As shown in Section 5.4, relaxing this assumption will likely lead to larger estimates of peer effects along this margin. Second, the functional form assumptions made in our structural approach may simplify estimation at the expense of some flexibility. Specifically, Equations (3) (project quality) and (4) (agent preferences) omit certain terms such that benefit and productivity shocks can be point identified. These assumptions may bias our parameter estimates away from their true values. Subsequent work would do well to relax these assumptions by either additional structure, data, or a more flexible estimation strategy.

6.6 Counterfactual Analysis

Value of Peer Effects

While the presence of positive[82] peer effects precludes a socially optimal level of contribution under private provision, they may increase equilibrium contribution beyond what would be provided in a world without peer influence. In this sense, peer effects have the potential to effectively subsidize the cost of private provision. Indeed, the preliminary analysis of the structural estimates in the preceding subsection gives reason to believe that peer behavior can drive preference and cost heterogeneity along both the extensive and intensive margins, albeit to differing degrees. To gauge the "value-added" by highly nuanced peer effects in terms of aggregate contribution labor, we use the estimates of the structural model to derive a counterfactual equilibrium in which peer effects are absent.

We consider the following policy counterfactual: suppose peer effects do not exist. In other words, past peer contribution does not influence an individual's likelihood of contribution and private benefit and productivity shocks are uncorrelated for individuals who decide to contribute. This scenario roughly corresponds to "siloed" development: agents independently contribute to a public good but do so without interaction with peers or the contribution levels of peers. What is the resulting level of contribution? Specifically, we begin by setting extensive and intensive margin peer effect parameters to zero: $\boldsymbol{\gamma} = \boldsymbol{\delta} = \mathbf{0}$. We then use the remaining parameter estimates $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, b_{pt})$

[82]Note that in the canonical public goods model of Samuelson (1954), negative peer effects (e.g., a congestion externality) could potentially offset the classic positive externality that drives free-riding and under-contribution relative to the social optimum.

to re-simulate the data generating process described by the structural model for the entire sample period.[83] To compare the relative impact of extensive and intensive margin effects, we also simulate a counterfactual under which extensive margin peer effects exist while only intensive margin peer effects are absent.

The results of these counterfactuals, in terms of aggregate contribution across all projects, are summarized alongside the observed data in Figure 15. Two key patterns emerge. First, the counterfactual without peer effects results in aggregate contribution approximately 55.6% lower compared with the observed equilibrium. By June 2019, aggregate contribution across all projects in the observed sample totals in excess of 5.519 million commits. Under the counterfactual scenario with no peer effects, aggregate contribution is reduced to approximately 2.452 million commits. If contributors make 2–3 commits per labor hour on average, a back-of-the-envelope calculation for this shortfall of 3.067 million commits implies a loss of 1–1.5 million OSS labor hours relative to the observed equilibrium. The median hourly wage of $52.95 for software developers in the U.S. suggests the value-added by OSS peer effects in our sample at $54.132 to $81.199 million USD. Second, extensive margin peer effects constitute the overwhelming of the value added. Figure 15 shows that the counterfactual scenario in which only extensive margin peer effects exist (i.e., $\boldsymbol{\delta} = 0$) closely matches the observed equilibrium under which both extensive and intensive margin effects are active. As discussed in Sections 5.4 and 6.5, the diminished role of intensive margin effects may be a result of our narrowly tailored definition for peer influence.

7 Discussion

Using the context of OSS, we have studied the influence of peer effects on the private provision of public goods in detail. We use both reduced form and structural approaches to (1) address non-random selection, (2) distinctly model intensive and extensive margin peer effects, and (3) disentangle marginal private benefits and costs of contribution as distinct channels of influence. Our findings in both approaches are consistent with anecdotal evidence: OSS project growth is largely driven by some combination of dedicated large-share core contributors and the arrival of many small-share contributors. We find little evidence that peers make each other more productive

[83] We use data $(a_{ipt}, y_{pt}, x_{it})$ and recovered shocks (v_{ipt}, c_{ipt}) to "seed" initial conditions for period $t =$ 2008–04–01.

on average: contemporaneous intensive margin peer effects are heterogeneous across projects but do appear to have been larger in the early days of GitHub. Moreover, structural estimates suggest the effect of peer influence on average is that agents contribute greater levels when their peers do, but at greater marginal cost. Our counterfactual analysis seeks to estimate the value-added of peer effects in terms of private public good provision. Driven almost exclusively by extensive margin peer effects, we find that cumulative contribution is approximately 56% lower under the scenario where peer effects are not present.

We can interpret the findings of this analysis to highlight some limitations for the potential for peer effects to foster the production of public information goods. We find that while extensive margin effects can drive a significant share of contribution, these effects are decreasing in the size of the peer group. This may arise either (1) if small share contributors free-ride on the efforts of dominant core contributors or (2) become less likely to contribute once a project matures in size. Moreover, the lack of strong, positive peer influence in contribution productivity along the intensive margin suggests that any strategic complementarity or substitution in contribution may simply offset on net. Compared with previous studies which document strong pro-social effects to collaboratively produced public goods (Zhang and Zhu, 2011; Slivko, 2014), peer effects in the production of more complex information goods like OSS may be significantly more nuanced.

As noted above, a key takeaway of this study is that agents differ in their willingness to contribute their labor towards the sustained maintenance of OSS used by a larger community. The extent to which peer effects matter for sustaining the quality of OSS public goods likely depends on both (1) the project's use valuation from the wider community and (2) the project's position as a component of OSS infrastructure (Eghbal, 2016).[84] Whereas the study of peer effects in this chapter focuses purely on the production side of public goods, a promising direction for future research is to explore the welfare implications for behavioral patterns uncovered thus far.[85] Better characterizations of optimal contribution patterns that consider the wider set of beneficiaries to

[84]For example, if OSS production fits a combinatoric production model in which developers make small, specialized contributions and move on to other projects, peer effects may be of less importance to delivering an efficient equilibrium. On the other hand, network externalities might amplify the importance of maintenance labor and positive peer effects. Consider the example of OpenSSL and the Heartbleed Bug. OSS infrastructure that is widely depended upon but maintained by a small group could likely benefit from additional contribution labor that can at least partially be generated through peer effects.

[85]In other words, subsequent studies would do well to distinguish between the welfare implications of production versus sustained maintenance for complex public information goods like OSS.

OSS quality allow the researcher to better discern the extent to which peer influences on collaboration truly matter. Such efforts can continue to place the economic significance of peer effects, externalities, and public good production into context for OSS.

References

Andersen-Gott, Morten, Gheorghita Ghinea, and Bendik Bygstad, "Why do commercial companies contribute to open source software?," *International journal of information management*, 2012, *32* (2), 106–117.

Andreoni, James, "Impure altruism and donations to public goods: A theory of warm-glow giving," *The economic journal*, 1990, *100* (401), 464–477.

Andrews, Isaiah, James H Stock, and Liyang Sun, "Weak instruments in instrumental variables regression: Theory and practice," *Annual Review of Economics*, 2019, *11*, 727–753.

Angrist, Joshua D, "The perils of peer effects," *Labour Economics*, 2014, *30*, 98–108.

_ **and Victor Lavy**, "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *The Quarterly journal of economics*, 1999, *114* (2), 533–575.

Archambault, Caroline, Matthieu Chemin, and Joost de Laat, "Can peers increase the voluntary contributions in community driven projects? Evidence from a field experiment," *Journal of Economic Behavior & Organization*, 2016, *132*, 62–77.

Arcidiacono, Peter and Sean Nicholson, "Peer effects in medical school," *Journal of public Economics*, 2005, *89* (2-3), 327–350.

Athey, Susan and Glenn Ellison, "Dynamics of open source movements," *Journal of Economics & Management Strategy*, 2014, *23* (2), 294–316.

Baldwin, Carliss Y and Kim B Clark, "The architecture of participation: Does code architecture mitigate free riding in the open source development model?," *Management science*, 2006, *52* (7), 1116–1127.

Benkler, Yochai, "Coase's Penguin, or, Linux and" The Nature of the Firm"," *Yale law journal*, 2002, pp. 369–446.

Bergquist, Magnus and Jan Ljungberg, "The power of gifts: organizing social relationships in open source communities," *Information Systems Journal*, 2001, *11* (4), 305–320.

Bergstrom, Theodore, Lawrence Blume, and Hal Varian, "On the private provision of public goods," *Journal of public economics*, 1986, *29* (1), 25–49.

Bessen, James, "Open source software: Free provision of complex public goods," in "The economics of open source software development," Elsevier, 2006, pp. 57–81.

Boehm, Barry, *Software Engineering Economics*, Vol. 197, Prentice-Hall, New York, 1981.

Bollinger, Bryan and Kenneth Gillingham, "Peer effects in the diffusion of solar photovoltaic panels," *Marketing Science*, 2012, *31* (6), 900–912.

Bonaccorsi, Andrea and Cristina Rossi Lamastra, "Altruistic individuals, selfish firms? The structure of motivation in Open Source software," *The Structure of Motivation in Open Source Software*, 2003.

_ **, Silvia Giannangeli, and Cristina Rossi**, "Entry strategies under competing standards: Hybrid business models in the open source software industry," *Management science*, 2006, *52* (7), 1085–1098.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin, "Identification of peer effects through social networks," *Journal of econometrics*, 2009, *150* (1), 41–55.

Brown, C. Titus, "A framework for thinking about Open Source Sustainability?," 7 2018. Accessed: 2021–12–04.

Carrell, Scott E, Bruce I Sacerdote, and James E West, "From natural variation to optimal policy? The Lucas critique meets peer effects," Technical Report, National Bureau of Economic Research 2011.

Ciliberto, Federico, Amalia R Miller, Helena Skyt Nielsen, and Marianne Simonsen, "Playing the fertility game at work: An equilibrium model of peer effects," *International Eco-*

nomic Review, 2016, *57* (3), 827–856.

Coase, Ronald Harry, "The nature of the firm," *economica*, 1937, *4* (16), 386–405.

Cornelissen, Thomas, Christian Dustmann, and Uta Schönberg, "Peer effects in the workplace," *American Economic Review*, 2017, *107* (2), 425–56.

Cornes, Richard and Todd Sandler, "The simple analytics of pure public good provision," *Economica*, 1985, *52* (205), 103–116.

Dahl, Gordon B, Katrine V Løken, and Magne Mogstad, "Peer effects in program participation," *American Economic Review*, 2014, *104* (7), 2049–74.

Eghbal, Nadia, *Roads and bridges: The unseen labor behind our digital infrastructure*, Ford Foundation, 2016.

_ , *Working in public: the making and maintenance of open source software*, Stripe Press, 2020.

Elliott, Matthew and Benjamin Golub, "A network approach to public goods," *Journal of Political Economy*, 2019, *127* (2), 730–776.

Falk, Armin and Andrea Ichino, "Clean evidence on peer effects," *Journal of labor economics*, 2006, *24* (1), 39–57.

Fehr, Ernst and Simon Gächter, "Cooperation and punishment in public goods experiments," *American Economic Review*, 2000, *90* (4), 980–994.

Fershtman, Chaim and Neil Gandal, "The determinants of output per contributor in open source projects: An empirical examination," *Available at SSRN 515282*, 2004.

_ **and** _ , "Open source software: Motivation and restrictive licensing," *International Economics and Economic Policy*, 2007, *4* (2), 209–225.

_ **and** _ , "Direct and indirect knowledge spillovers: the "social network" of open-source projects," *The RAND Journal of Economics*, 2011, *42* (1), 70–91.

Fischbacher, Urs and Simon Gächter, "Heterogeneous social preferences and the dynamics of free riding in public goods," 2006.

Fogel, Karl, *Producing open source software: How to run a successful free software project*, " O'Reilly Media, Inc.", 2005.

Giorgi, Giacomo De, Michele Pellizzari, and Silvia Redaelli, "Identification of social interactions through partially overlapping peer groups," *American Economic Journal: Applied Economics*, 2010, *2* (2), 241–75.

GitHub, Inc., "The State of the Octoverse," 2020. Accessed: 2021–08–27.

_ , "Choose an open source license: Licenses," 2022. Accessed: 2022–06–16.

Glaeser, Edward L, Bruce I Sacerdote, and Jose A Scheinkman, "The social multiplier," *Journal of the European Economic Association*, 2003, *1* (2-3), 345–353.

Glazer, Amihai and Kai A Konrad, "A signaling explanation for charity," *The American Economic Review*, 1996, *86* (4), 1019–1028.

Goldsmith-Pinkham, Paul and Guido W Imbens, "Social networks and the identification of peer effects," *Journal of Business & Economic Statistics*, 2013, *31* (3), 253–264.

Gousios, Georgios, "The GHTorrent dataset and tool suite," in "Proceedings of the 10th Working Conference on Mining Software Repositories" MSR '13 IEEE Press Piscataway, NJ, USA 2013, pp. 233–236.

Graham, Bryan S, "Methods of identification in social networks," *Annu. Rev. Econ.*, 2015, *7* (1), 465–485.

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

Guryan, Jonathan, Kory Kroft, and Matthew J Notowidigdo, "Peer effects in the workplace: Evidence from random groupings in professional golf tournaments," *American Economic Journal: Applied Economics*, 2009, *1* (4), 34–68.

Hahn, Jungpil, Jae Yun Moon, and Chen Zhang, "Emergence of new project teams from open source software developer networks: Impact of prior collaboration ties," *Information Systems Research*, 2008, *19* (3), 369–391.

Heckman, James J, "Sample selection bias as a specification error," *Econometrica: Journal of the econometric society*, 1979, pp. 153–161.

Hilton, Michael, Timothy Tunnell, Kai Huang, Darko Marinov, and Danny Dig, "Usage, costs, and benefits of continuous integration in open-source projects," in "2016 31st IEEE/ACM International Conference on Automated Software Engineering (ASE)" 2016, pp. 426–437.

Holländer, Heinz, "A social exchange approach to voluntary cooperation," *The American Economic Review*, 1990, pp. 1157–1167.

Hsieh, Chih-Sheng, Michael D Konig, Xiaodong Liu, and Christian Zimmermann, "Superstar Economists: Coauthorship networks and research output," *Available at SSRN 3266432*, 2018.

_ , _ , _ , and _ , "Collaboration in bipartite networks, with an application to coauthorship networks," 2020.

Jacobsen, Mark, Jacob LaRiviere, and Michael Price, "Public policy and the private provision of public goods under heterogeneous preferences," *Journal of the Association of Environmental and Resource Economists*, 2017, *4* (1), 243–280.

Johnson, Justin Pappas, "Open source software: Private provision of a public good," *Journal of Economics & Management Strategy*, 2002, *11* (4), 637–662.

Jr, Frederick P Brooks, *The mythical man-month: essays on software engineering*, Pearson Education, 1995.

Kerner, Sean Michael, "Heartbleed SSL Flaw's True Cost Will Take Time to Tally," 4 2014. Accessed: 2021–12–04.

Kotchen, Matthew J, "Voluntary provision of public goods for bads: A theory of environmental offsets," *The Economic Journal*, 2009, *119* (537), 883–899.

Krueger, Alan B, "Economic considerations and class size," *The economic journal*, 2003, *113* (485), F34–F63.

Lakhani, Karim R and Robert G Wolf, "Why hackers do what they do: Understanding motivation and effort in free/open source software projects," *Open Source Software Projects (September 2003)*, 2003.

Lee, Lung-Fei, "Identification and estimation of econometric models with group interactions, contextual factors and fixed effects," *Journal of Econometrics*, 2007, *140* (2), 333–374.

Lerner, Josh and Jean Tirole, "Some simple economics of open source," *The journal of industrial economics*, 2002, *50* (2), 197–234.

_ and _ , "The scope of open source licensing," *Journal of Law, Economics, and Organization*, 2005, *21* (1), 20–56.

Lewbel, Arthur, "The identification zoo: Meanings of identification in econometrics," *Journal of Economic Literature*, 2019, *57* (4), 835–903.

Lindquist, Matthew J, Jan Sauermann, and Yves Zenou, "Network effects on worker productivity," 2015.

Manski, Charles F, "Identification of endogenous social effects: The reflection problem," *The review of economic studies*, 1993, *60* (3), 531–542.

Mas, Alexandre and Enrico Moretti, "Peers at work," *American Economic Review*, 2009, *99* (1), 112–45.

McQuaid, Mike, "The Open Source Contributor Funnel (or: Why People Don't Contribute To Your Open Source Project)," 8 2018. Accessed: 2021–12–04.

Meneely, Andrew, Alberto C Rodriguez Tejeda, Brian Spates, Shannon Trudeau,

Danielle Neuberger, Katherine Whitlock, Christopher Ketant, and Kayla Davis, "An empirical investigation of socio-technical code review metrics and security vulnerabilities," in "Proceedings of the 6th International Workshop on Social Software Engineering" 2014, pp. 37–44.

Munaiah, Nuthan, Steven Kroh, Craig Cabrey, and Meiyappan Nagappan, "Curating github for engineered software projects," *Empirical Software Engineering*, 2017, *22* (6), 3219–3253.

Nagle, Frank, "Open source software and firm productivity," *Management Science*, 2019, *65* (3), 1191–1215.

_ , James Dana, Jennifer Hoffman, Steven Randazzo, and Yanuo Zhou, "Census II of Free and Open Source Software — Application Libraries," Technical Report, The Linux Foundation and The Laboratory for Innovation Science at Harvard March 2022.

Nitzan, Shmuel and Richard E Romano, "Private provision of a discrete public good with uncertain cost," *Journal of Public Economics*, 1990, *42* (3), 357–370.

Oberhaus, Daniel, "The Complicated Economy of Open Source Software," February 2019. Accessed: 2021–09-23.

Olea, José Luis Montiel and Carolin Pflueger, "A robust test for weak instruments," *Journal of Business & Economic Statistics*, 2013, *31* (3), 358–369.

Open Source Initiative, "The Open Source Definition," 3 2007. Accessed: 2022–06–16.

Ostrom, Elinor, *Governing the commons: The evolution of institutions for collective action*, Cambridge university press, 1990.

Patnam, Manasa, "Corporate networks and peer effects in firm policies," in "Emerging Markets Finance Conference, Indira Gandhi Institute of Development Research" 2011.

Raymond, Eric, "The cathedral and the bazaar," *Knowledge, Technology & Policy*, 1999, *12* (3), 23–49.

Roberts, Jeffrey A, Il-Horn Hann, and Sandra A Slaughter, "Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the Apache projects," *Management science*, 2006, *52* (7), 984–999.

Sacerdote, Bruce, "Peer effects with random assignment: Results for Dartmouth roommates," *The Quarterly journal of economics*, 2001, *116* (2), 681–704.

Samuelson, Paul A, "The pure theory of public expenditure," *The review of economics and statistics*, 1954, *36* (4), 387–389.

Sherwood, Paul, "Estimating the costs of open-source development," 2015. Embedded Linux Conference Europe.

Sholler, Dan, Igor Steinmacher, Denae Ford, Mara Averick, Mike Hoye, and Greg Wilson, "Ten simple rules for helping newcomers become contributors to open projects," *PLoS computational biology*, 2019, *15* (9), e1007296.

Slivko, Olga, "Peer effects in collaborative content generation: The evidence from German Wikipedia," *ZEW-Centre for European Economic Research Discussion Paper*, 2014, (14-128).

Sorhus, Sindre, "(@sindresorhus) Some observations from having merged thousands of pull requests in the past few years," 5 2019. Accessed: 2021–12–04.

_ , "Sindre Sorhus (personal web page)," 12 2021. Accessed: 2021–12–04.

Stiglitz, Joseph E, "Public goods in open economies with heterogeneous individuals," 1981.

_ , "The theory of local public goods twenty-five years after Tiebout: A perspective," Technical Report, National Bureau of Economic Research 1982.

The OpenSSL Project Authors, "OpenSSL," 2021. Copyright 1999–2021.

Tiebout, Charles M, "A pure theory of local expenditures," *Journal of political economy*, 1956, *64* (5), 416–424.

U.S. Bureau of Labor Statistics, "Software Developers, Quality Assurance Analysts, and Testers : Occupational Outlook Handbook," Sep 2021.

Wickham, Hadley, *R packages: organize, test, document, and share your code*, " O'Reilly Media, Inc.", 2015.

Zhang, Xiaoquan Michael and Feng Zhu, "Group size and incentives to contribute: A natural experiment at Chinese Wikipedia," *American Economic Review*, 2011, *101* (4), 1601–15.

Zhao, Jun, Hea-Jung Kim, and Hyoung-Moon Kim, "New EM-type algorithms for the Heckman selection model," *Computational Statistics & Data Analysis*, 2020, *146*, 106930.

A Tables

Table 1: Descriptive Statistics – Primary Measures in Empirical Sample

Measure	Notation	Count	Mean	SD	Min	Median	Max
Project commits (total)	a_p	2,287	2,490	7,720	23	825	188,292
Individual commits (total)	a_i	107,921	53	669	1	2	186,464
Project commits (monthly)	a_{pt}	96,453	59	294	1	16	73,161
Individual commits (monthly)	a_{it}	421,879	13	129	1	3	73,145
Cumulative individual commits (monthly)	$\tilde{a}_{it}$	421,879	278	1,109	1	28	186,464
Cumulative project commits (monthly)	$\tilde{a}_{pt}$	96,453	1,989	6,295	1	532	188,292
Individual commits (project-month)	a_{ipt}	440,111	13	126	1	3	73,145
Peer commits (project-month)	$a_{\text{-}ipt}$	440,111	188	398	0	59	73,160
Cumulative individual commits (project-month)	$\tilde{a}_{ipt}$	440,111	256	1,076	1	23	186,447
Cumulative peer commits (project-month)	$\tilde{a}_{\text{-}ipt}$	440,111	2,096	6,630	0	262	124,932
Number of peers (project-month)	n_{ipt}	440,111	17	29	0	7	310
Cumulative GitHub Stars (project-month)	y_{pt}	96,294	910	2,924	0	161	81,817
GitHub active days (monthly)	gad_{it}	411,427	3.88	4.67	1	2	31

Table 2: Reduced Form – Individual Level Peer Effects Estimates (Baseline Estimates for peer effect δ from Equation (1))

	OLS Individual Commits			IV 2SLS Individual Commits		
	(1)	(2)	(3)	(4)	(5)	(6)
Peer Commits	0.0078	0.0065	0.0035	-0.0035	0.0089	0.0102
	(0.0011)	(0.0018)	(0.0034)	(0.0015)	(0.0037)	(0.0251)
Individual Commits (cumulative)	-	0.0332	0.0529	-	0.0333	0.0529
	-	(0.0134)	(0.0434)	-	(0.0134)	(0.0435)
Individual Commits (previous month)	-	0.2604	0.1853	-	0.2604	0.1853
	-	(0.1763)	(0.0902)	-	(0.1763)	(0.0902)
Peer Commits (cumulative)	-	-0.0020	-0.0024	-	-0.0020	-0.0024
	-	(0.0009)	(0.0019)	-	(0.0009)	(0.0019)
Peer Commits (previous month)	-	0.0051	0.0036	-	0.0044	0.0039
	-	(0.0022)	(0.0032)	-	(0.0024)	(0.0046)
Peer Group Size	-	0.0017	0.0898	-	-0.0093	0.0158
	-	(0.0662)	(0.1457)	-	(0.0553)	(0.3760)
Controls	No	Yes	Yes	No	Yes	Yes
Fixed Effects	No	No	Yes	No	No	Yes
N	440,111	433,867	433,867	436,287	433,867	433,867
R^2	0.0006	0.1802	0.2268	-0.0007	0.1802	0.2267
First stage F statistic				6,520	1,151	64.37

Note: Columns (1)–(6) present the coefficient estimate $\hat{\delta}$ from Equation (1) in which aggregate peer commits are regressed on individual commits. Standard errors appear in parentheses below the coefficient estimate. Columns (1), (2), (4), and (5) use heteroskedasticity-robust standard errors while Columns (3) and (6) cluster standard errors by project. Column (4) through (6) additionally report the cluster-robust F-statistic from the first stage of the two-stage least squares procedure. Control variables include three lags of individual and peer commits, cumulative individual and peers commits, project quality measured in GitHub stars, quadratic terms for project age and peer group size, and dummy variables indicating if the individual is a project owner, project member, or if they are affiliated with a firm. Fixed effects included are individual, project, and year-month.

Table 3: Reduced Form – Individual Level Peer Effects (Estimates of peer effect δ from Equation (1) with covariate interaction terms)

	OLS			IV 2SLS		
	(1)	(2)	(3)	(4)	(5)	(6)
Peer Commits	0.0078	0.0065	0.0167	-0.0035	0.0497	-0.0721
	(0.0011)	(0.0020)	(0.0105)	(0.0015)	(0.0354)	(0.1562)
Peer Commits ×		-1.47 $\times 10^{-5}$	-0.0002		-0.0002	0.0003
Peer Group Size		(2.59 $\times 10^{-5}$)	(0.0001)		(0.0001)	(0.0009)
Peer Commits ×		0.0004	0.0004		0.0004	0.0004
Lagged Individual Commits		(0.0001)	(0.0002)		(0.0001)	(0.0002)
Peer Commits ×		-3.09 $\times 10^{-6}$	-2.32 $\times 10^{-6}$		-2.54 $\times 10^{-6}$	-2.6 $\times 10^{-6}$
Lagged Peer Commits		(1.54 $\times 10^{-6}$)	(1.13 $\times 10^{-6}$)		(1.16 $\times 10^{-6}$)	(1.5 $\times 10^{-6}$)
Peer Commits ×		-3.82 $\times 10^{-5}$	-4.9 $\times 10^{-5}$		-3.74 $\times 10^{-5}$	-4.89 $\times 10^{-5}$
Cumulative Individual Commits		(1.86 $\times 10^{-5}$)	(3.39 $\times 10^{-5}$)		(1.81 $\times 10^{-5}$)	(3.36 $\times 10^{-5}$)
Peer Commits ×		1.81 $\times 10^{-6}$	1.93 $\times 10^{-6}$		1.83 $\times 10^{-6}$	2.03 $\times 10^{-6}$
Cumulative Peer Commits		(1.07 $\times 10^{-6}$)	(6.3 $\times 10^{-7}$)		(1.08 $\times 10^{-6}$)	(7.57 $\times 10^{-7}$)
Peer Commits ×		6.18 $\times 10^{-8}$	2.82 $\times 10^{-7}$		5.24 $\times 10^{-8}$	-3.03 $\times 10^{-7}$
Project Quality		(1.87 $\times 10^{-8}$)	(3.5 $\times 10^{-7}$)		(2.42 $\times 10^{-8}$)	(9.98 $\times 10^{-7}$)
Peer Commits ×		0.0451	0.3513		0.0329	0.3732
Project Owner		(0.0753)	(0.2380)		(0.0848)	(0.2455)
Peer Commits ×		0.0093	0.0070		0.0001	0.0311
Project Member		(0.0065)	(0.0066)		(0.0039)	(0.0412)
Peer Commits ×		-3.55 $\times 10^{-6}$	7.89 $\times 10^{-6}$		-1.23 $\times 10^{-5}$	3.17 $\times 10^{-5}$
Project Age		(1.54 $\times 10^{-6}$)	(7.89 $\times 10^{-6}$)		(8.59 $\times 10^{-6}$)	(4.73 $\times 10^{-5}$)
Controls	No	Yes	Yes	No	Yes	Yes
Fixed Effects	No	No	Yes	No	No	Yes
N	440,111	433,867	433,867	436,287	433,867	433,867
R^2	0.0006	0.1910	0.2372	-0.0007	0.1891	0.2351
First stage F statistic				6,520	389.5	57.177

Note: Columns (1)–(6) estimate specifications corresponding to those presented in Table 2 with the inclusion of terms interacted with aggregate peer effort (RHS endogenous term).

Table 4: Reduced Form – Temporal Heterogeneity in Individual Level Peer Effects (Estimates of peer effect δ from Equation (1) for subsamples disaggregated by time period)

	OLS			IV 2SLS		
	(1)	(2)	(3)	(4)	(5)	(6)
2008 - 2012	0.0119	0.0271	0.0267	-0.0104	-0.0270	-0.0359
	(0.0011)	(0.0017)	(0.0077)	(0.0020)	(0.0235)	(0.0487)
N	20,399	20,209	20,209	20,262	20,262	20,209
R^2	0.0077	0.5158	0.6288	-0.0196	0.4807	0.6010
First stage F statistic				3,677	57.00	85.29
2012 - 2016	0.0178	0.0155	0.0228	-0.0237	0.0037	-0.1262
	(0.0005)	(0.0006)	(0.0028)	(0.0008)	(0.0016)	(0.2851)
N	146,778	145,952	145,952	146,256	145,952	145,952
R^2	0.0279	0.4975	0.5837	-0.1244	0.4937	0.3716
First stage F statistic				6,637.1	1,543	8.405
2016 - 2019	0.0056	0.0052	0.0032	-0.0013	0.0165	-0.0009
	(0.0010)	(0.0017)	(0.0026)	(0.0020)	(0.0076)	(0.0385)
N	272,934	267,706	267,706	269,769	267,706	269,706
R^2	0.0003	0.1835	0.2696	-0.0001	0.1829	0.2696
First stage F statistic				3,491.0	607.9	35.63

Note: Columns (1)–(6) estimate specifications corresponding to those presented in Table 2 distinctly for sub-samples disaggregated by time period.

Table 5: Reduced Form – Beyond Contemporaneous Individual Level Peer Effect Estimates (Estimates for peer effect δ from Equation (13))

	OLS Individual Commits			IV 2SLS Individual Commits		
	(1)	(2)	(3)	(4)	(5)	(6)
Peer Commits	0.0139	0.0212	0.0068	-0.0145	0.0575	0.0881
	(0.0009)	(0.0042)	(0.0070)	(0.0018)	(0.0110)	(0.0914)
Individual Commits (cumulative)	-	-0.0051	-0.0063	-	-0.0074	-0.077
	-	(0.0020)	(0.0047)	-	(0.0021)	(0.0056)
Individual Commits (previous month)	-	0.3158	0.1036	-	0.3014	0.0949
	-	(0.2402)	(0.2433)	-	(0.2355)	(0.2410)
Peer Group Size	-	-0.0175	-0.0071	-	-0.7290	-1.835
	-	(0.1270)	(0.4223)	-	(0.2148)	(2.068)
Controls	No	Yes	Yes	No	Yes	Yes
Fixed Effects	No	No	Yes	No	No	Yes
N	440,111	433,867	433,867	436,287	433,867	433,867
R^2	0.0021	0.1802	0.2274	-0.0066	0.2200	0.3100
First stage F statistic				4,376	124.9	32.16

Note: Columns (1)–(6) present the coefficient estimate $\hat{\delta}$ from Equation (13) in which aggregate peer commits from the preceding 3 months are regressed on individual commits for the subsequent 3 months. Covariate controls and fixed effects correspond to the estimates in Table 2.

Table 6: Reduced Form – Project-Level Estimates (Historical Project Contribution regressed on Contemporaneous Project Contribution)

	OLS Project Commits						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Project Commits (1 month prior)	0.4502	0.3188	0.3188	0.3186	0.2804	0.3186	0.2803
	(0.2216)	(0.1998)	(0.2000)	(0.0387)	(0.0449)	(0.0386)	(0.0450)
Project Commits (2 months prior)	-	-0.0383	-0.0383	-0.0384	-0.0632	-0.0384	-0.0633
	-	(0.0666)	(0.0659)	(0.0289)	(0.0349)	(0.0288)	(0.0350)
Project Commits (3 months prior)	-	0.1224	0.1223	0.1222	0.0882	0.1221	0.0880
	-	(0.0391)	(0.0384)	(0.0311)	(0.0394)	(0.0309)	(0.0396)
Project Commits (cumulative)	-	0.0156	0.0156	0.0157	0.0159	0.0157	0.0160
	-	(0.0039)	(0.0033)	(0.0067)	(0.0148)	(0.0067)	(0.0149)
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effects: Time	No	No	Yes	No	No	Yes	Yes
Fixed Effects: Project	No	No	No	No	Yes	No	Yes
Fixed Effects: Language	No	No	No	Yes	No	Yes	No
N	96,453	96,294	96,294	96,294	96,294	96,294	96,294
R^2	0.19555	0.27487	0.27567	0.27494	0.30079	0.27575	0.30159

Note: Columns (1)–(7) contain coefficient estimates for current month total project contribution regressed on project contribution in previous months and the cumulative project contribution. Other controls include lagged and cumulative numbers of project contributors, project quality, and quadratic terms for project age. Columns (1) and (2) report heteroskedasticity-robust standard errors, Column (3) clusters standard errors by month, Columns (4) and (6) by project language, and Columns (5) and (7) by project.

Table 7: Reduced Form – Project-Level Estimates (Historical Number of Contributors regressed on Contemporaneous Number of Contributors)

	OLS						
	Number of Project Contributors						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Number of contributors (1 month prior)	0.8822	0.6082	0.6085	0.6080	0.5187	0.6082	0.5188
	(0.0166)	(0.0430)	(0.0459)	(0.0422)	(0.0380)	(0.0423)	(0.0382)
Number of contributors (2 months prior)	-	0.1175	0.1167	0.1173	0.0778	0.1166	0.0774
	-	(0.0408)	(0.0477)	(0.0248)	(0.0346)	(0.0251)	(0.0348)
Number of contributors (3 months prior)	-	0.1772	0.1769	0.1769	0.1307	0.1767	0.1306
	-	(0.0311)	(0.0336)	(0.0266)	(0.0196)	(0.0263)	(0.0194)
Number of contributors (cumulative)	-	0.0007	0.0007	0.0007	-0.0008	0.0007	-0.0008
	-	(0.0002)	(0.0002)	(0.0002)	(0.0004)	(0.0003)	(0.0008)
Project Commits (cumulative)	-	1.01e-5	1.04e-5	1.01e-5	0.0001	1.05e-5	0.0001
	-	(8.09e-6)	(8.07e-6)	(8.1e-6)	(1.69e-5)	(1.13e-5)	(4.33e-5)
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effects: Time	No	No	Yes	No	No	Yes	Yes
Fixed Effects: Project	No	No	No	No	Yes	No	Yes
Fixed Effects: Language	No	No	No	Yes	No	Yes	No
N	96,453	96,294	96,294	96,294	96,294	96,294	96,294
R^2	0.77107	0.79349	0.79494	0.79354	0.81159	0.79499	0.81299

Note: Columns (1)–(7) contain coefficient estimates for the number of contributors for a project in the current month regressed on the number of contributors in previous months and the cumulative number of project contributors. Other controls include lagged and cumulative project contribution, project quality, and quadratic terms for project age. Columns (1) and (2) report heteroskedasticity-robust standard errors, Column (3) clusters standard errors by month, Columns (4) and (6) by project language, and Columns (5) and (7) by project.

Table 8: Structural Model Notation

$i, j \in \mathcal{N}$	agents where $\lvert\mathcal{N}\rvert = N$
$p \in \mathcal{P}$	OSS projects where $\lvert\mathcal{P}\rvert = P$
$t \in \mathcal{T}$	time periods where $\lvert\mathcal{T}\rvert = T$
$a_{ipt} \in \mathbb{R}_+$	agent i's contribution to project p in period t
$y_{pt} \in \mathbb{R}$	quality of project p in t
$b_{pt} = \frac{\partial y_{pt}}{\partial a_{ipt}}$	marginal product of labor in terms of public good quality
$x_{it} \in \mathbb{R}_+$	agent i's consumption of numeraire good (e.g. time)
$\omega_{it} \in \mathbb{R}_+$	agent i's numeraire endowment
z_{ipt}	agent i's latent "ability" in project p at time t (see Equation (6))
z_p	project p's latent "ability threshold" (see Equation (6))
v_{ipt}	private contribution benefit for agent i in project p at time t
c_{ipt}	contribution cost (inverse productivity) for agent i in project p at time t
$\boldsymbol{\gamma}$	extensive margin peer effects (see Equation (10))
δ_v	intensive margin peer effects for marginal private benefits of contribution
δ_c	intensive margin peer effects for marginal private costs of contribution (see Equation (11))
$\boldsymbol{\beta}_z$	control variable parameters for latent agent productivity z_{ipt} in extensive margin decision (see Equation (10))
$\boldsymbol{\beta}_v$	control variable parameters for marginal private benefits of contribution (see Equation (12))
$\boldsymbol{\beta}_c$	control variable parameters for marginal private cost of contribution (see Equation (12))
ϵ^z_{ipt}	Unobserved factors influencing extensive margin decision (see Equation (10))
ϵ^v_{ipt}	Unobserved factors influencing marginal private benefit shock v_{ipt} (see Equation (12))
ϵ^c_{ipt}	Unobserved factors influencing marginal private cost shock c_{ipt} (see Equation (11))

B Figures

Figure 2: Example GitHub Repository Page – `twbs/bootstrap`

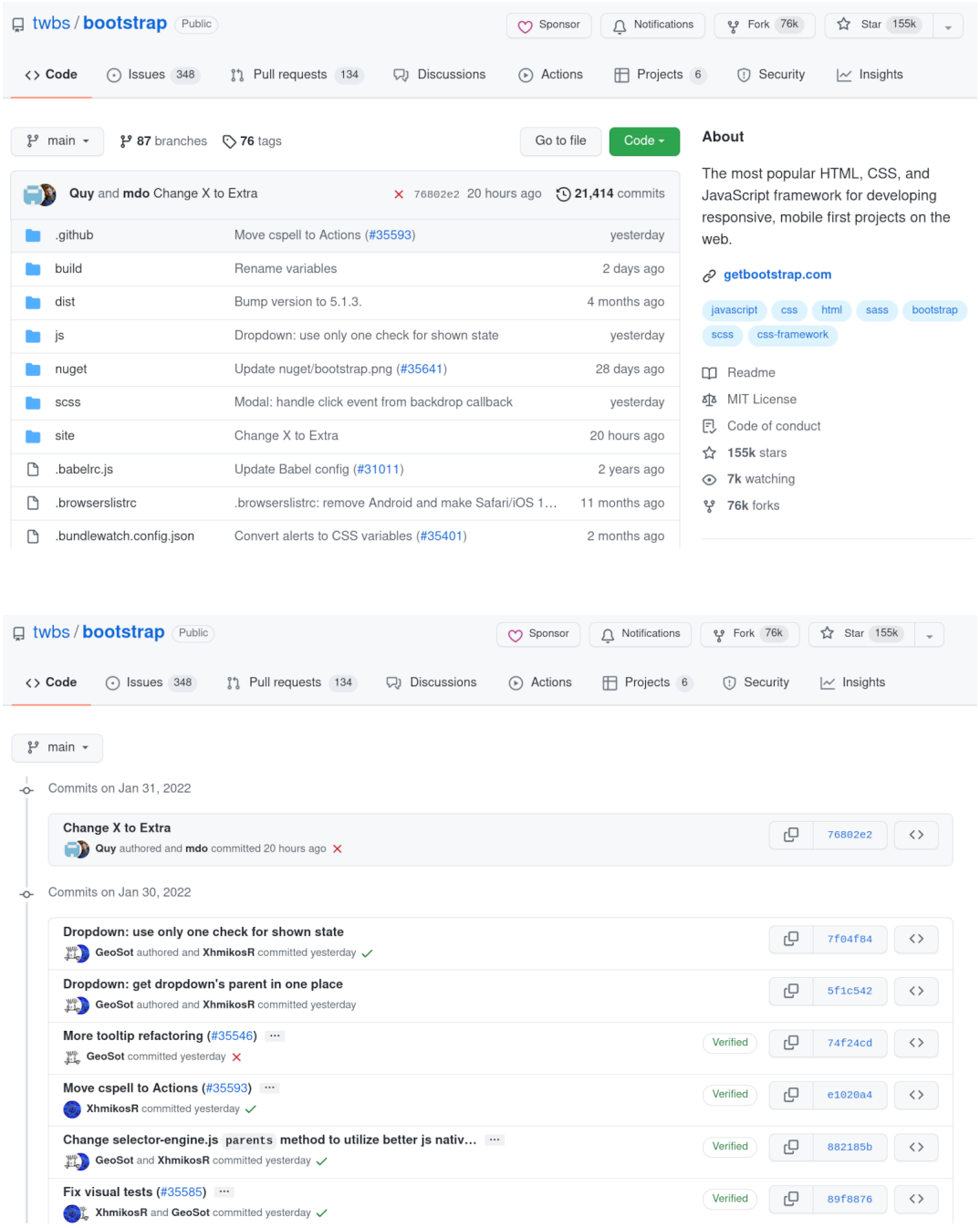


Figure 3: Descriptive Statistics – Project Creation Dates and Earliest Commits

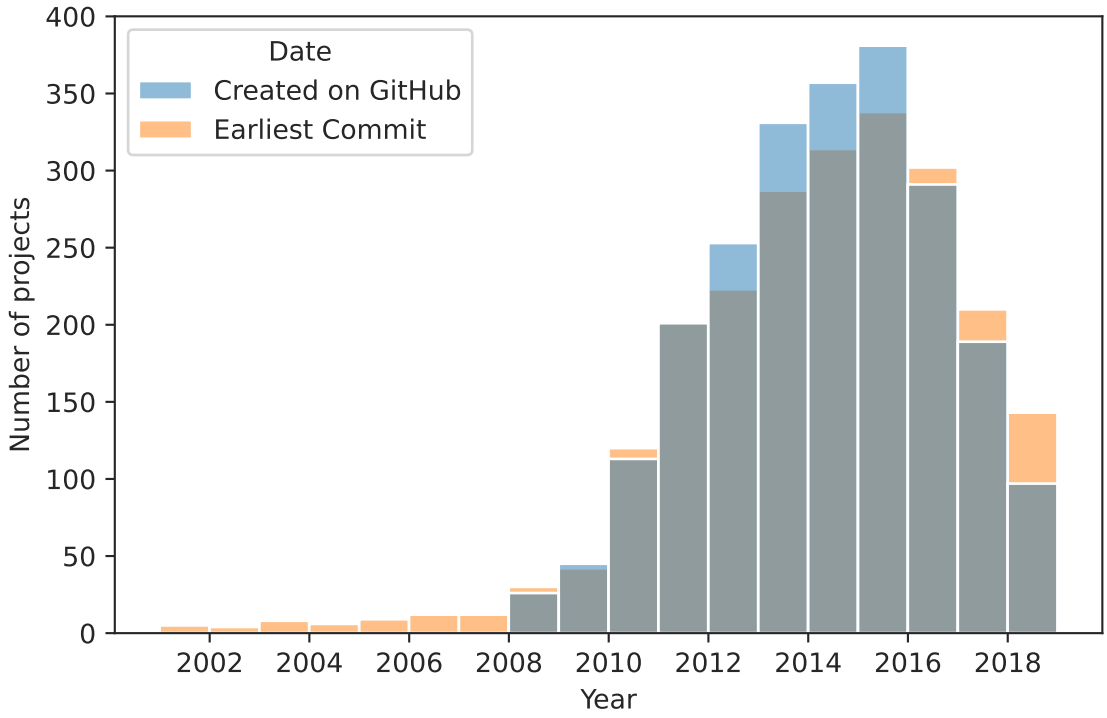


Figure 4: Descriptive Statistics – Distribution of Project-level Contribution Shares

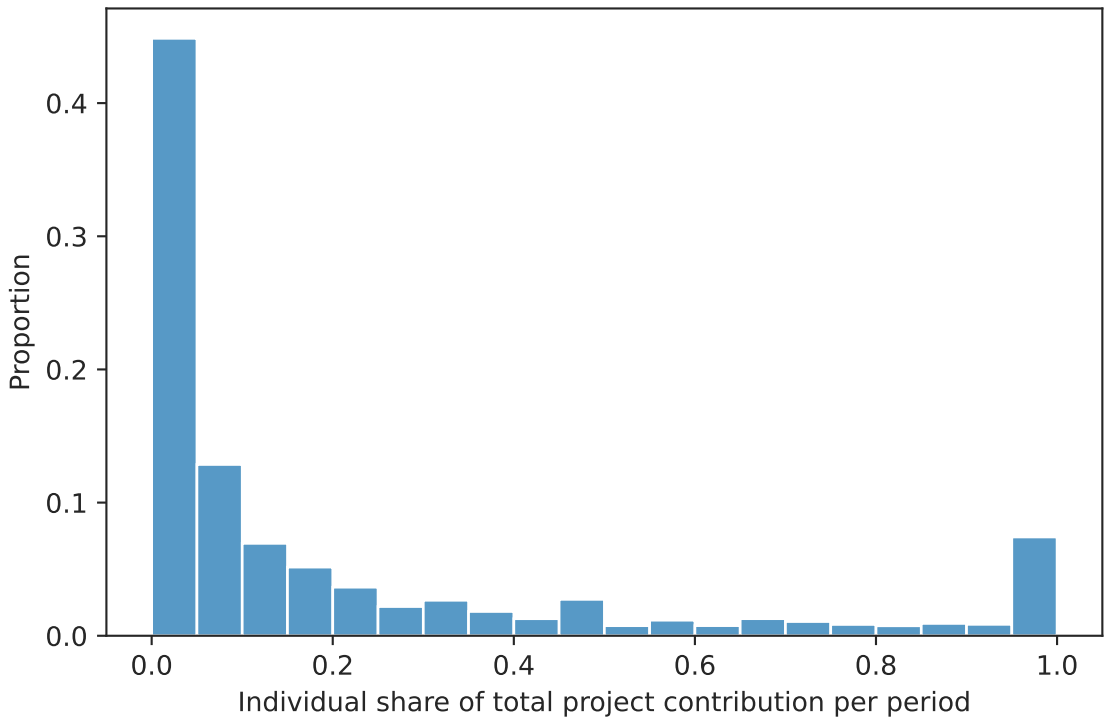


Figure 5: Descriptive Statistics – Aggregate contribution in sample

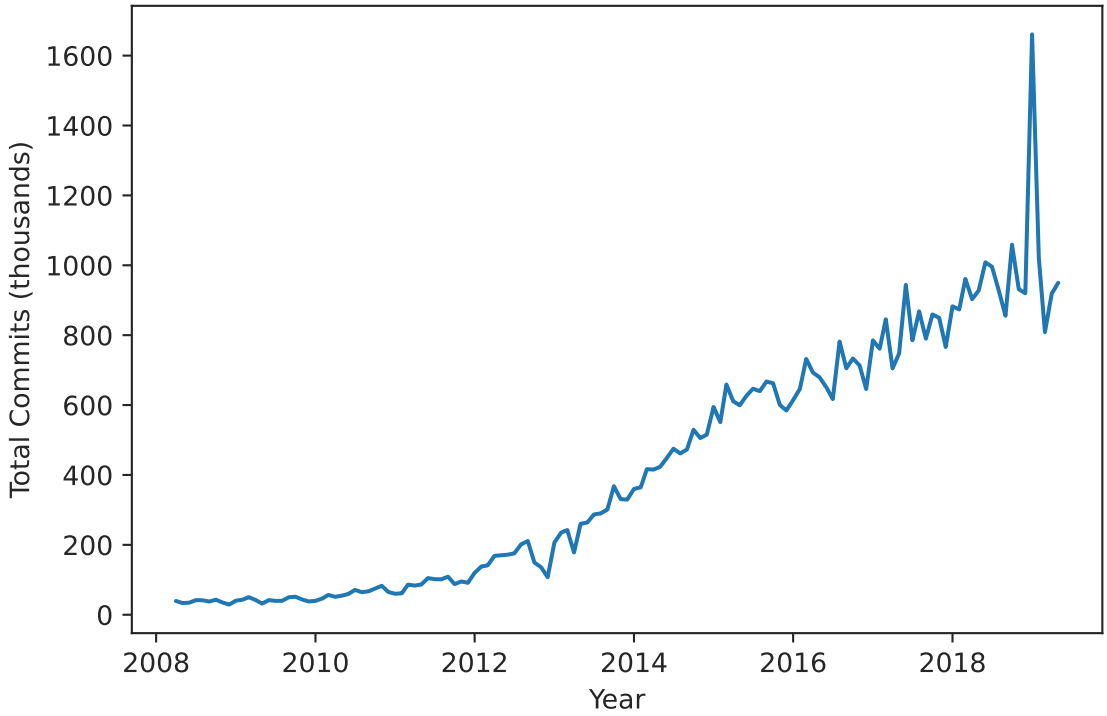


Figure 6: Descriptive Statistics – Distinct contributors in sample

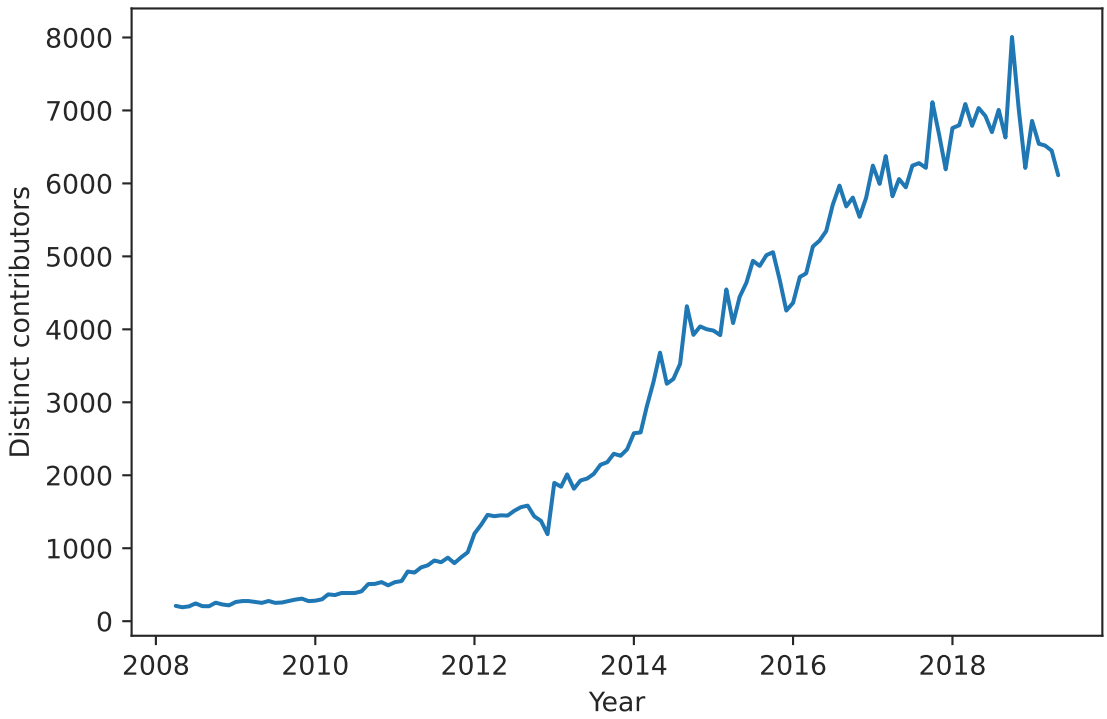


Figure 7: Descriptive Statistics – Mean individual and peer contribution per project

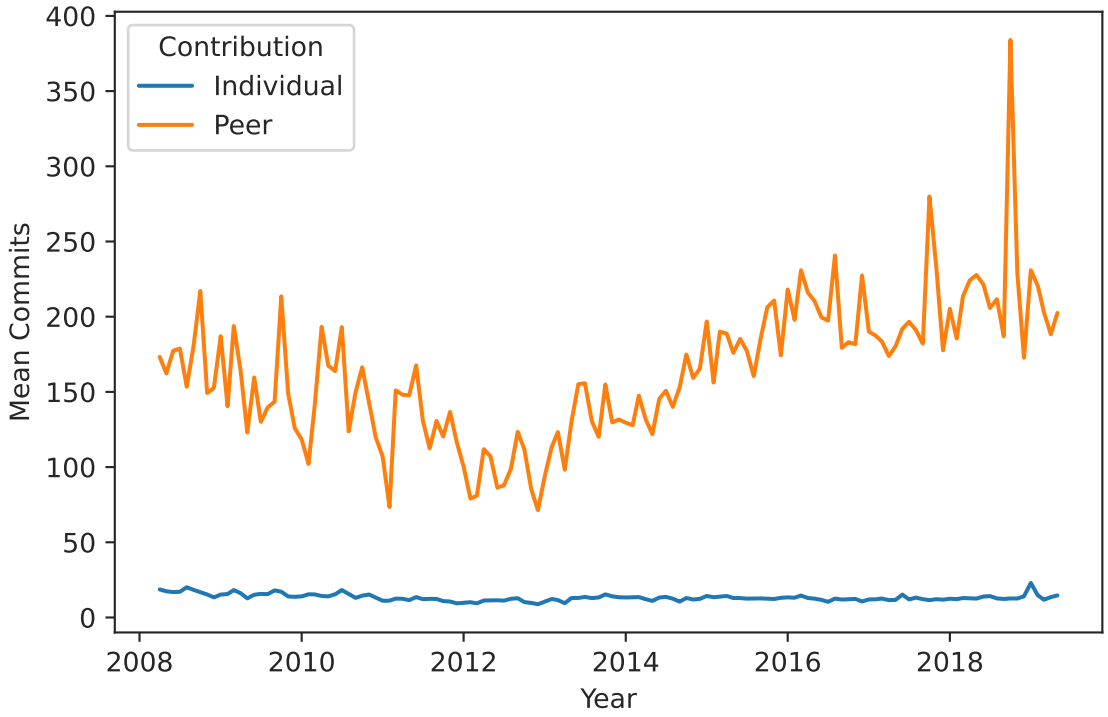


Figure 8: Reduced Form – Project Heterogeneity (Distribution of project-level estimates of $\hat{\delta}$ for Equation (1))

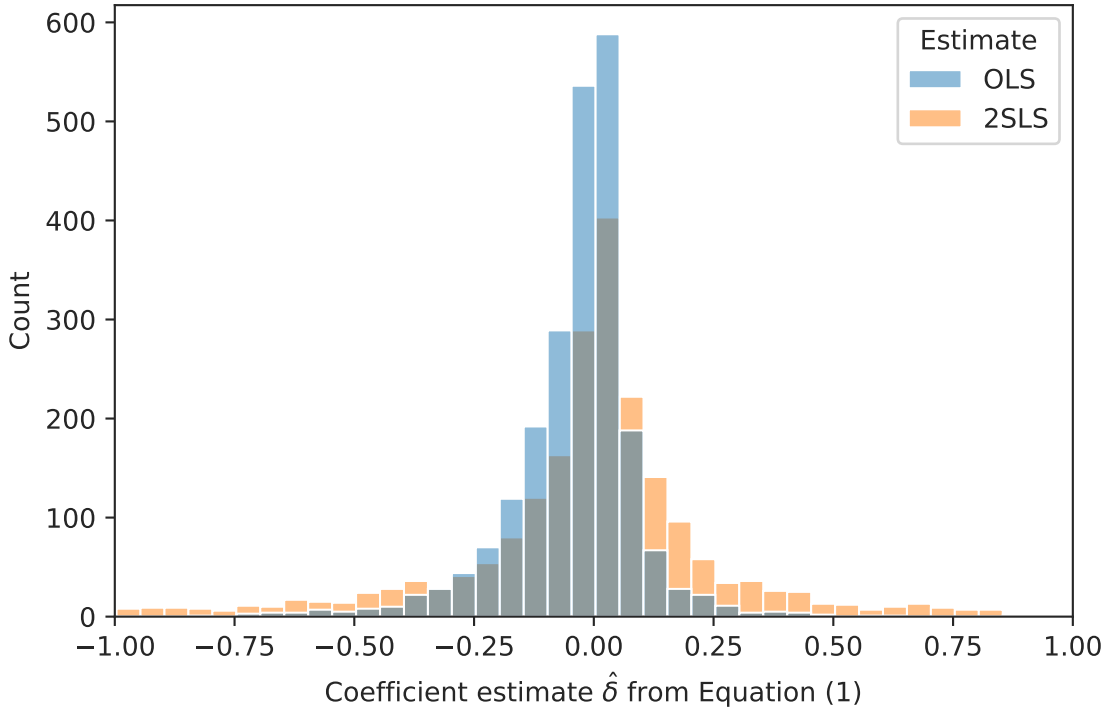


Figure 9: Reduced Form – Temporal Heterogeneity (Estimates for Equation (1) over sample period. The top subplot includes estimates for the peer effect coefficient $\hat{\delta}$ of Equation (1) within annual cross-sectional sub-samples. The bottom subplot includes estimates for the peer effect coefficient within cumulative sub-samples (i.e., all observations $\leq t$).)

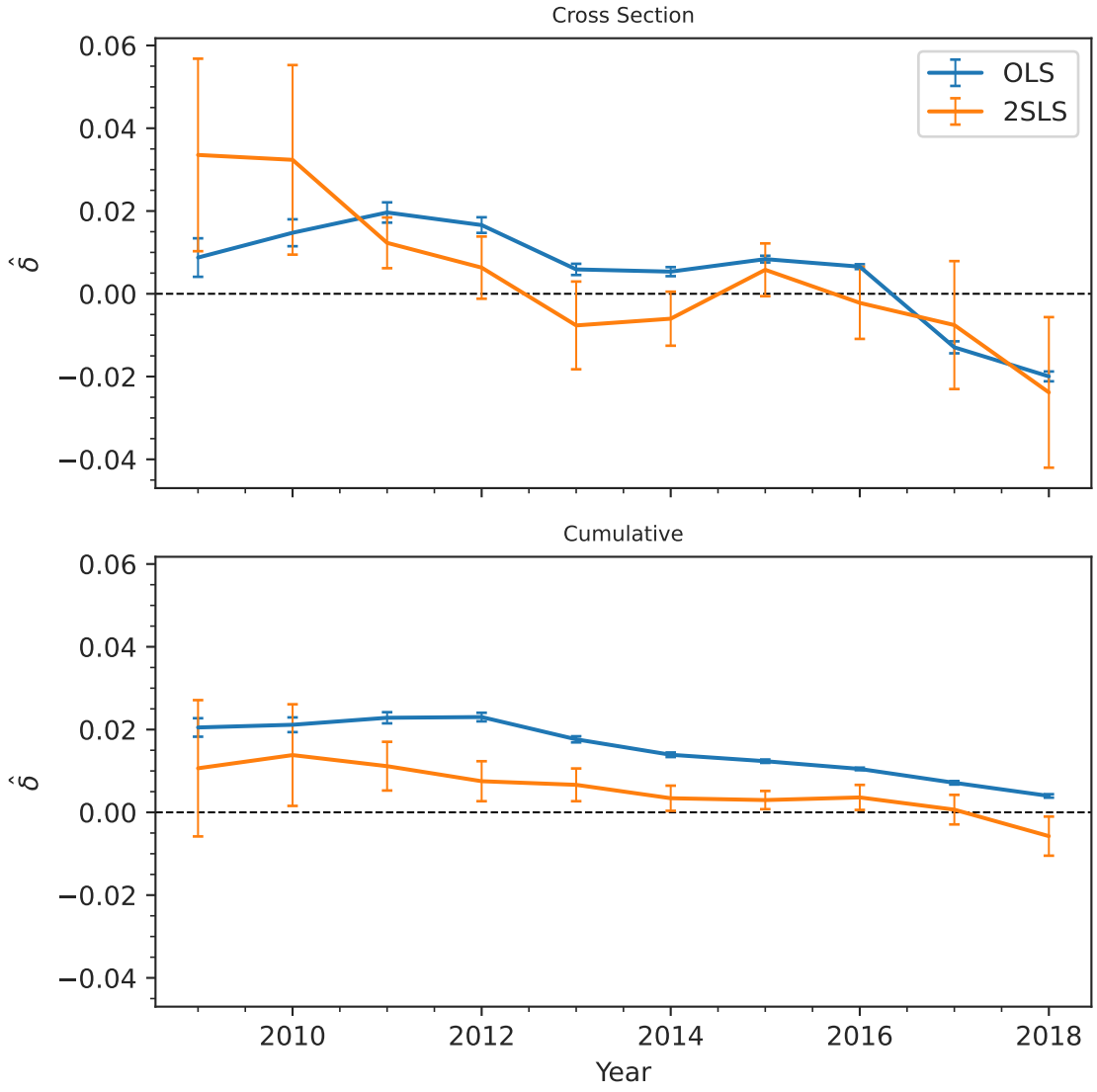


Figure 10: Reduced Form – Insider Contribution and Crowding Out (Estimates for δ in Equation (14))

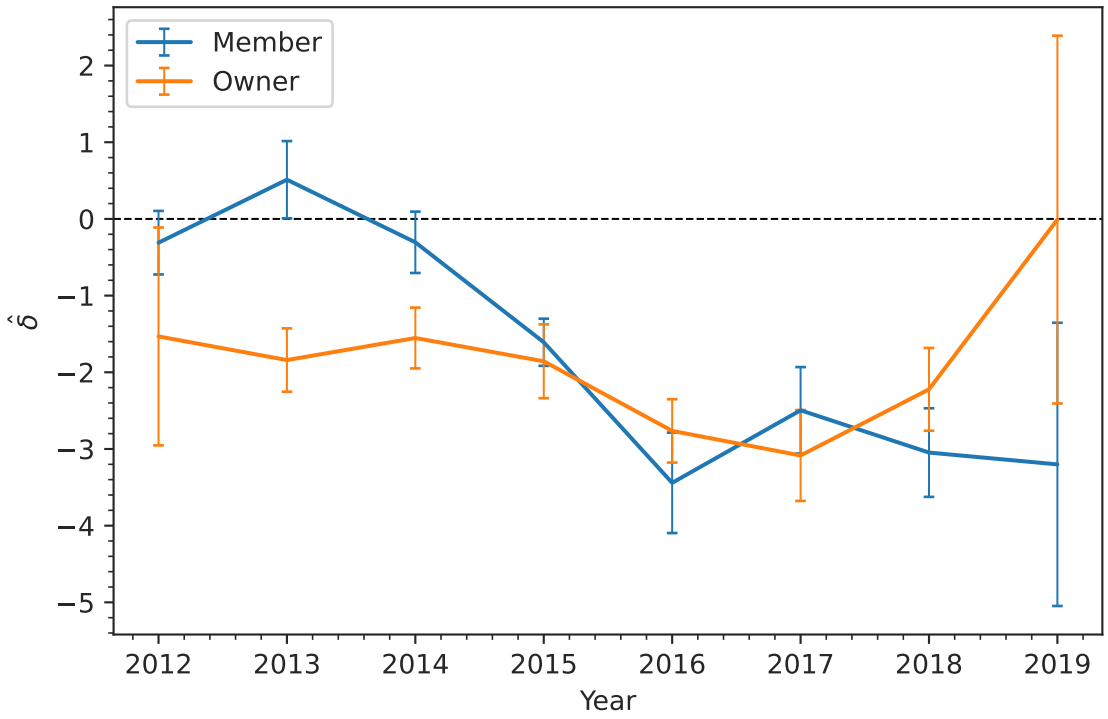


Figure 11: Structural Model – Recovered Benefit and Productivity Shocks v_{ipt}, c_{ipt} and marginal product of labor parameters b_{pt} for all observed contribution $a^{\star}_{ipt} > 0$ (Equation (8))

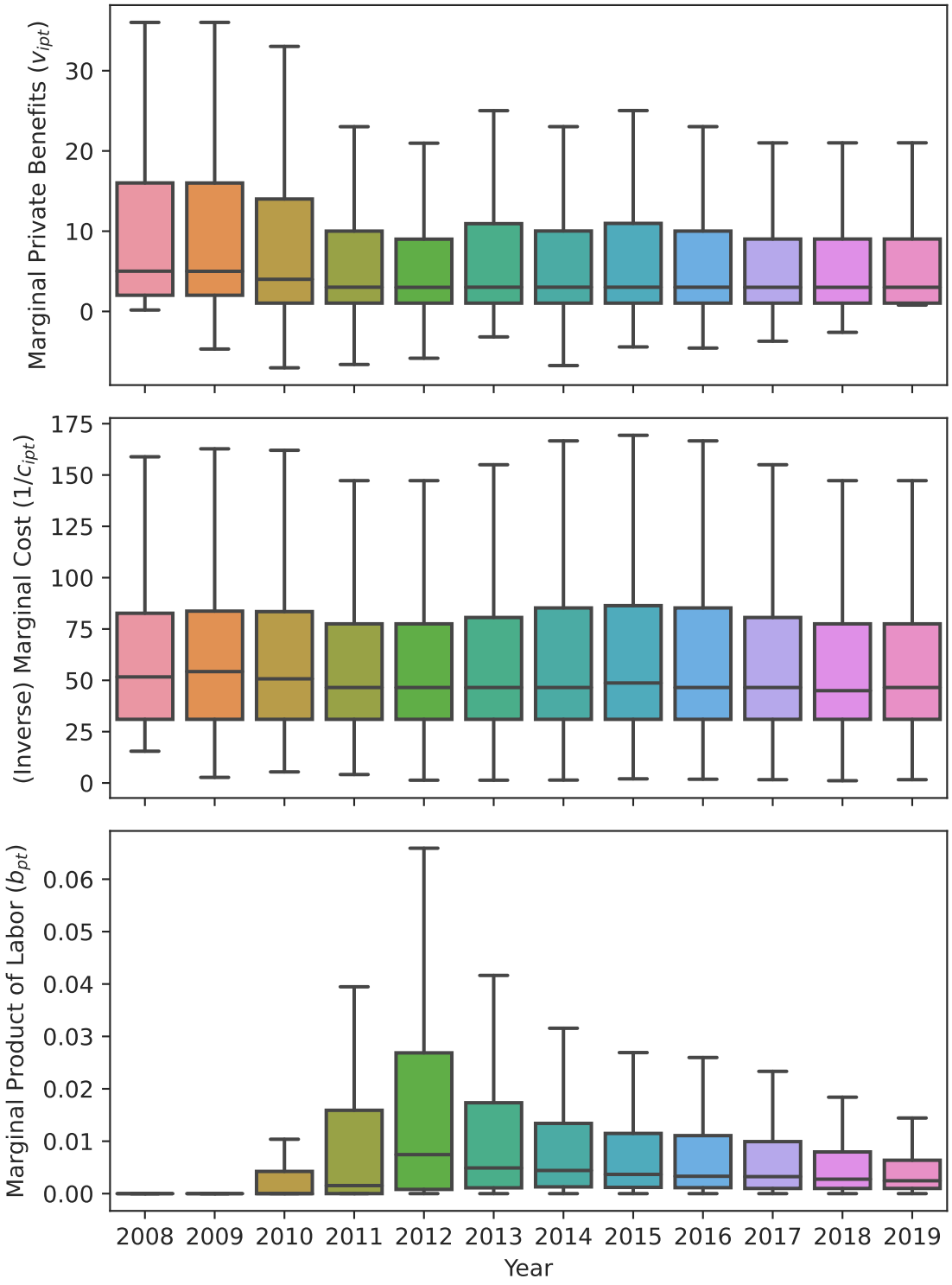


Figure 12: Structural Model – Correlation between Benefit and Productivity Shocks v_{ipt}, c_{ipt} over sample period

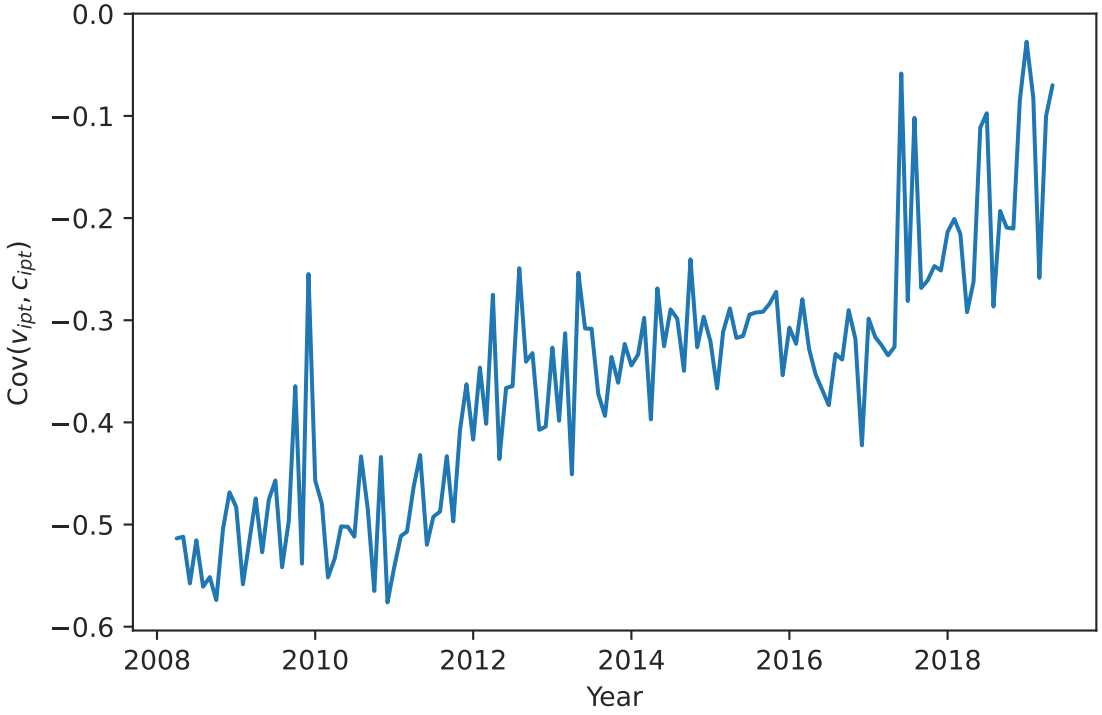


Figure 13: Structural Model – Extensive Margin Peer Effects (Project-level estimates for γ from Equation (10))

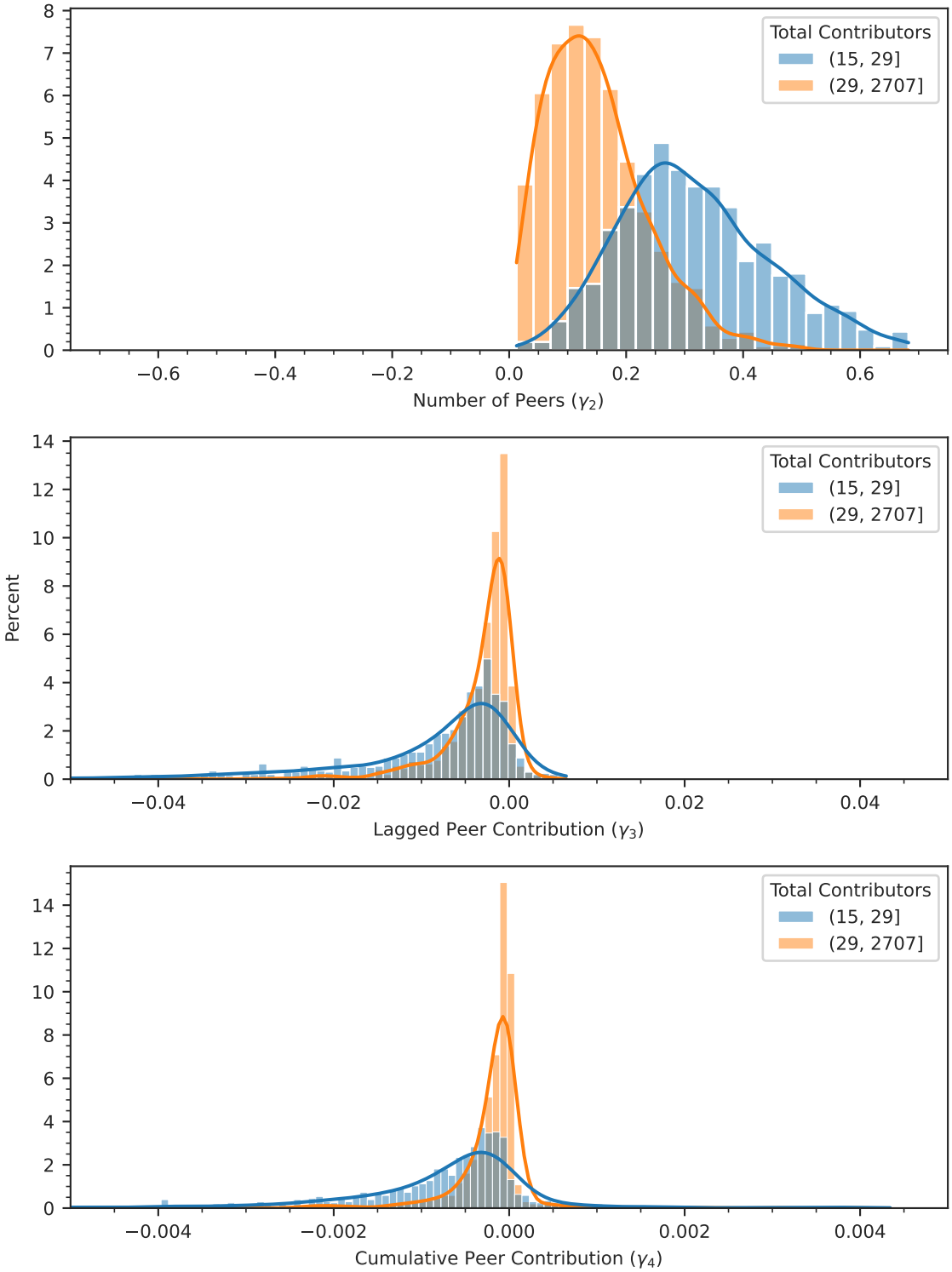


Figure 14: Structural Model – Intensive Margin Peer Effects (Project-level estimates for δ_v from Equation (12) and δ_c from Equation (11))

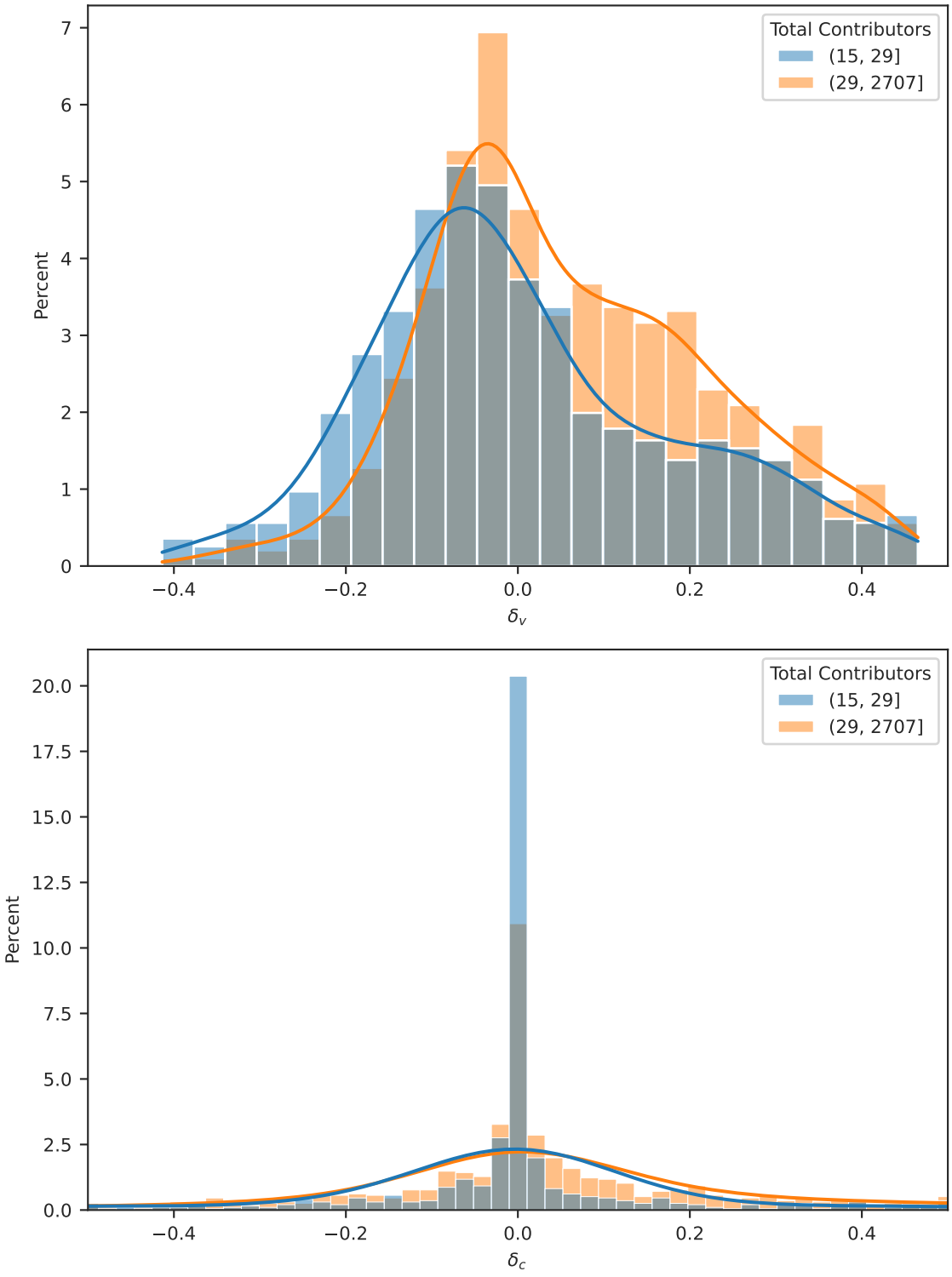
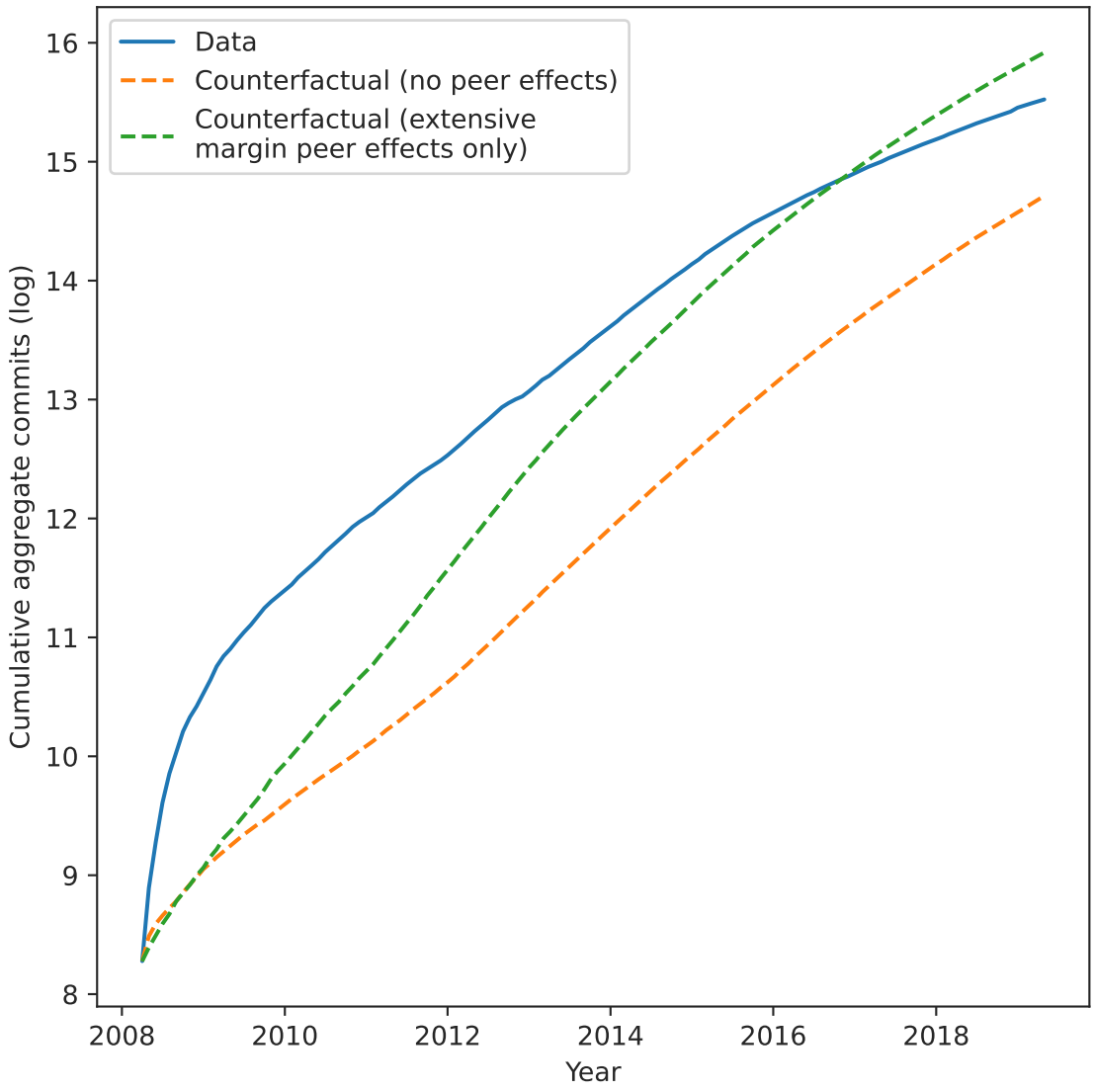


Figure 15: Structural Model – Counterfactual Growth in Aggregate Contribution without Peer Influence



C Data Details

Sources

We use several data sources for our empirical sample

- GHTorrent[86], an archive that seeks to provide an offline, historical record of all public activity on the GitHub platform. The data is very large (the June 2019 archive is 104 GB compressed) but includes scripts so that it can be loaded into a relational database management system for out-of-core analysis. As an alternative, the data is also hosted on Google Big Query.
- Project source code hosted on the GitHub platform. Projects are usually managed by a version control system (VCS) that, among many other technical features, records a chronological history of changes to the project's codebase. This allows us to create measures for project characteristics over each point in time in the project's history. On GitHub, the VCS tool used is `git`.

Sample Selection

As the number of projects recorded in the GHTorrent dataset is rather unwieldy for analysis by conventional means, we resort to sampling. We use the following procedure to develop a sample of popular, collaborative OSS projects hosted on GitHub:

1. From the set of public GitHub projects created before 2019–06–01, select the subset with (1) 15 or more distinct contributors and (2) 100 cumulative "stars". Denote this set of top projects $\mathfrak{P}$.
2. Take a 10% random sample $\mathcal{P} \subset \mathfrak{P}$ from the set of top projects. This set of core project will form the basis of projects considered in both the reduced form and structural analysis.
3. For all projects $p \in \mathcal{P}$, determine the set of agents $\mathcal{N}_p \equiv \{i \mid a_{ipt} > 0 \, \forall t \in \mathcal{T}\}$ that contribute to p. For core projects $\mathcal{P}$, contribution is observed over time periods $t \in \mathcal{T}$ where $\inf(\mathcal{T}) =$ 2009–11–01 and $\sup(\mathcal{T}) =$ 2019–05–01. Collect all core agents into the set $\mathcal{N} \equiv \cup_{p \in \mathcal{P}} \mathcal{N}_p$.
4. With $\mathcal{N}, \mathcal{P}$ and $\mathcal{T}$ defined, we can proceed in collecting measures of contribution levels, project characteristics, and agent characteristics.

[86] Source: https://ghtorrent.org/

D Additional Reduced Form Results

We provide some deeper analysis into the reduced form peer effects estimates in an effort to (1) provide robust support for the baseline peer effects estimates in Table 2 and (2) disentangle the various forces embedded in the full sample estimates.

Interactions

Various observable factors may be associated with different levels of peer effect on contribution. For example, agents larger or higher quality projects may respond differently to the contribution levels of their peers. Peer effects might also vary by the size of the project or peer group itself. We investigate these effects by estimating a version of Equation (1) that includes interactions between peer effort and various observables. We present interaction term coefficient estimates in Table 3.

At first glance, it is apparent that the interactions between peer effort and these various factors are second order compared to the primary peer effect. Moreover, most are statistically in consequential at conventional levels. It is interesting to note, however, that peer effects are strongest when an agent has some cumulative history of contribution with the project. This effect is stronger than the influence of peer group size and project quality. We interpret this effect as evidence to the notion that agents form strong affinities to OSS projects and contribute to them with limited concern over other exogenous factors. There is weak evidence that peer effects are stronger for agents invested in the project, such as owners and members, but these effects are not statistically different from zero across specifications.

Temporal Heterogeneity

Given the dramatic growth of OSS participation on GitHub, it is likely that peer effects in the early days of the platform are different from later years. Table 4 collects estimates of the specifications in Table 2 for different time periods. Two patterns emerge. First, positive peer effects are stronger in earlier periods. The Column (3) OLS estimates for years 2008 through 2012 are 0.0208 and statistically different from zero at conventional significance levels (compared with -0.0014 for the full sample). This estimate falls to -0.0088 for years 2016 through 2019. It should be noted that the number of observations in this subsamples are 20,209 and 267,706 respectively. Second, comparing Columns (3) and (6) across time periods in Table 4 suggests some evidence that the OLS estimates are positively biased. Finally, we further disaggregate the sample by time to estimate Equation (1) for (1) annual cross-sectional sub-samples and (2) cumulative sub-samples (i.e., for all observations $\leq t$ for $t \in \{2008, 2009, \ldots, 2019\}$). We plot these coefficient estimates in Figure 9.

Both Table 2 and Figure 9 suggest that peer effects were more likely to be positive in the early days of OSS on GitHub. This is consistent with Eghbal (2020)'s observation that "platforms broke the commons": early OSS collaboration likely featured smaller, more cohesive project communities in which work was distributed evenly. As GitHub grew in size, the arrival of many, small-share contributors helped grow projects in aggregate but coincide with diminished estimates for the peer

effect.

Project Heterogeneity

There is also considerable project-level heterogeneity in peer effects. Figure 8 plots the distribution of peer effects obtained by estimating Equation (1) for each project individually. A key takeaway from Figure 8 is that after accounting for covariates, peer effect estimates are surprisingly rather symmetric around the null hypothesis of $\delta = 0$. The share of projects in which peer and individual effort are substitutes and those in which they are complements is relatively well-balanced within the sample.

Beyond Contemporaneous Peer Effects

The contemporaneous specification in Equation (1) is likely too narrowly defined to capture peer effects that develop over a span longer than a single month. Since OSS contribution is public record, peer effects in a general sense need not be strictly contemporaneous. We estimate a version of this specification that seeks to estimate the effect of recent peer contribution (e.g., previous three months) on subsequent individual contribution (e.g. preceding three months):

$$\sum_{\tau=0}^{2} a_{ipt+\tau} = \delta \sum_{\tau=0}^{2} a_{\text{-}ipt-\tau} + \boldsymbol{\beta}' \boldsymbol{X}_{ipt} + \epsilon_{ipt}. \tag{13}$$

We present estimates of the specification above in Table 5. The estimates in Table 5 are larger in magnitude compared to the baseline results in Table 2, suggesting peer effects are stronger when considered under a wider temporal bandwidth.

Project Level Effects

To begin to see how contribution patterns manifest along the extensive margin, we aggregate contribution to the project-month level and regress (1) aggregate project contribution on cumulative and lagged contribution (Table 6) and (2) the number of contributors on cumulative and lagged contributor groups (Table 7). The estimates in columns (6) and (7) of Table 6 imply that aggregate project contribution is greater, on average, when lagged project contribution is greater. Similarly, columns (6) and (7) in Table 7 demonstrate that contributor peer groups is autocorrelated on average. Both of these results suggest that past contribution behavior predicts future participation along the extensive margin. We explore the potential for extensive margin peer influence more thoroughly in Section 6.3.

Insider Contribution and Crowding Out

An alternative way to look at peer groups within OSS projects is to distinguish between project "insiders" and contributors from the wider community. A natural question is whether contribution from project insiders crowds out contributions from project outsiders. The wider community may

have strong incentives to free-ride on disproportionate contributions from dominant core contributors. We define a project insider as an individual who is either the nominal project owner or a member of the project. We aggregate individual contribution to the project level and split it into insider contribution a_{pt}^{in} and outsider contribution a_{pt}^{out}. Our "crowding-out" specification is a regression of outsider contribution on insider contribution and project level controls:

$$a_{pt}^{\text{out}} = \delta a_{pt}^{\text{in}} + \boldsymbol{\beta}' \boldsymbol{X}_{pt} + \epsilon_{pt}. \tag{14}$$

We estimate Equation (14) for each period and plot the coefficient estimates for $\hat{\delta}$ in Figure 10. Estimates for $\hat{\delta}$ are consistently negative and statistically significant, giving strong evidence for crowding out by project insiders. It is also worthy to note that crowding out appears to increase in later periods, coinciding with diminishing peer effects over time in Figure 9.

E Structural Estimation Details

Given data $(a_{ipt}, y_{pt}, x_{it})$ for all $i \in \mathcal{N}, p \in \mathcal{P}$, and $t \in \mathcal{T}$, we develop an estimation strategy to recover

1. Marginal product of labor parameters $\boldsymbol{b} = (b_{pt})$ from the project quality function in Equation (3).
2. Private benefit and productivity shocks $\boldsymbol{s} = (v_{ipt}, c_{ipt})$ for all $a^{\star}_{ipt} > 0$ from the equilibrium contribution level in Equation (8).
3. (Extensive margin peer effects) Parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta}_z)$ from Equation (10).
4. (Intensive margin peer effects) Parameters $(\delta_c, \delta_v, \boldsymbol{\beta}_c, \boldsymbol{\beta}_v)$ from Equations (11) and (12).

The parameters of interest are $\boldsymbol{\delta} = (\delta_c, \delta_v)$, which drive intensive margin peer effects, and $\boldsymbol{\gamma}$, which drive extensive margin peer effects. For each project $p \in \mathcal{P}$, our estimation strategy is as follows:

1. Assume disturbances are jointly normally distributed $(\epsilon^z_{ipt}, \epsilon^v_{ipt}, \epsilon^c_{ipt}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, independent and identically distributed between agents and time. Within the variance-covariance matrix $\boldsymbol{\Sigma}$, assume that $\sigma^2_z = 1$. This implies

$$\begin{bmatrix} \epsilon^z_{ipt} \\ \epsilon^v_{ipt} \\ \epsilon^c_{ipt} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{zv} & \sigma_{zc} \\ \sigma_{zv} & \sigma^2_v & \sigma_{vc} \\ \sigma_{zc} & \sigma_{vc} & \sigma^2_c \end{bmatrix} \right)$$

   Notice also that $\sigma_{zv} = \rho_{zv}\sigma_v$ and $\sigma_{zc} = \rho_{zc}\sigma_c$.
2. Given data (a_{ipt}, y_{pt}), recover $\boldsymbol{b}$ using Equation (3).
3. Given data $(a_{ipt}, y_{pt}, x_{it})$ and $\boldsymbol{b}$, recover shocks $\boldsymbol{s}$ using Equation (9), Equation (5), Equation (3) by means of GMM. Let $\mathcal{P}_{it} \equiv \{p \in \mathcal{P} \mid a^{\star}_{ipt} > 0\}$ be the subset of projects i contributes to in time t and $|\mathcal{P}_{it}| = P_{it}$. For each i and t, there are $2P_{it}$ unknowns: v_{ipt} and c_{ipt} for each $a^{\star}_{ipt} > 0$. There are P_{it} first order conditions from Equation (9), P_{it} equations for project quality form Equation (3), and one budget constraint (Equation (5)):

$$\begin{aligned} a_{ipt} &= b_{pt} + v_{ipt} - c_{ipt} > 0 \quad \forall p \in \mathcal{P}_{it} \\ y_{pt} &\leq b_{pt} \sum_j \sum_{s \leq t} a_{jps} \qquad \forall p \in \mathcal{P}_{it} \\ x_{it} &+ \sum_p c_{ipt} a_{ipt} \leq 1 \end{aligned} \tag{15}$$

   Combining the moment conditions in (15), the GMM formulation to recover (v_{ipt}, c_{ipt}) for

each agent i and period t given data $(a_{ipt}, y_{pt}, x_{it})$ and parameters b_{pt}, becomes

$$
\begin{aligned}
(v_{ipt}, c_{ipt}) = \underset{v_{ipt}, c_{ipt}}{\arg\min} \quad & \frac{1}{P_{it}} \sum_{p \in \mathcal{P}_{it}} (a_{ipt} - b_{pt} - v_{ipt} + c_{ipt})^2 \\
\text{s.t.} \quad & 0 < b_{pt} + v_{ipt} - c_{ipt} && \forall p \in \mathcal{P}_{it} \\
& 0 < c_{ipt} \leq 1 && \forall p \in \mathcal{P}_{it} \\
& y_{pt} \leq b_{pt} \sum_j \sum_t (b_{pt} + v_{jpt} - c_{jpt}) + b_{pt} \sum_j \sum_{s<t} a_{jps} && \forall p \in \mathcal{P}_{it} \\
& x_{it} + \sum_p c_{ipt}(b_{pt} + v_{ipt} - c_{ipt}) \leq 1
\end{aligned}
\tag{16}
$$

Across all agents and time periods, this implies $\sum_i \sum_t (2P_{it} + 1)$ total moment conditions and $\sum_i \sum_t 2P_{it}$ unknowns.

4. Given data $(d_{ipt} = 1\{a_{ipt} > 0\}, \boldsymbol{W}_{ipt}, \boldsymbol{X}_{ipt})$ and shocks $\boldsymbol{s}$ recover $(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, where $\boldsymbol{\delta} = (\delta_v, \delta_c)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_z, \boldsymbol{\beta}_v, \boldsymbol{\beta}_c)$, using the maximum likelihood estimation (MLE) framework for the Heckman Selection model described by Zhao et al. (2020). Collect quantities either observed as data or recovered in the previous stages of estimation into a vector $\boldsymbol{D} = (b_{pt}, d_{ipt}, v_{ipt}, c_{ipt}, \boldsymbol{W}_{ipt}, \boldsymbol{X}_{ipt})$. Collect remaining unknown parameters in a vector $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$. For each project $p \in \mathcal{P}$, the MLE optimization problem becomes

$$
\begin{aligned}
\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \quad & L(\boldsymbol{\theta} \mid \boldsymbol{D}) = \prod_i \prod_t \{f(v_{ipt} - c_{ipt} \mid d_{ipt} = 1) \Pr(d_{ipt} = 1)\}^{d_{ipt}} \Pr(d_{ipt} = 1)^{1 - d_{ipt}} \\
\text{s.t} \quad & f(v_{ipt} - c_{ipt} \mid d_{ipt} = 1) = \frac{1}{\sigma} \phi \left(\frac{\epsilon^v_{ipt} - \epsilon^c_{ipt}}{\sigma} \right) \frac{\Phi \left(\frac{\rho}{\sqrt{1-\rho^2}} \left(\frac{\epsilon^v_{ipt} - \epsilon^c_{ipt}}{\sigma} \right) + \frac{\boldsymbol{\gamma}' \boldsymbol{W}_{ipt} + \boldsymbol{\beta}_z' \boldsymbol{X}_{ipt}}{\sqrt{1-\rho^2}} \right)}{\Phi(\boldsymbol{\gamma}' \boldsymbol{W}_{ipt} + \boldsymbol{\beta}_z' \boldsymbol{X}_{ipt})} \\
& \Pr(d_{ipt} = d) = \Phi(\boldsymbol{\gamma}' \boldsymbol{W}_{ipt} + \boldsymbol{\beta}_z' \boldsymbol{X}_{ipt})^{d_{ipt}} \Phi(-\boldsymbol{\gamma}' \boldsymbol{W}_{ipt} - \boldsymbol{\beta}_z' \boldsymbol{X}_{ipt})^{1 - d_{ipt}} \\
& v_{ipt} = \delta_v \overline{v}_{\text{-}ipt} + \boldsymbol{\beta}_v' \boldsymbol{X}_{ipt} + \epsilon^v_{ipt} \\
& c_{ipt} = \delta_c \overline{c}_{\text{-}ipt} + \boldsymbol{\beta}_c' \boldsymbol{X}_{ipt} + \epsilon^c_{ipt} \\
& \sigma^2 = \sigma_v^2 + \sigma_c^2 - 2\rho_{vc}\sigma_v\sigma_c \\
& \rho\sigma = \sigma_{zv} - \sigma_{zc}
\end{aligned}
\tag{17}
$$

where ϕ and Φ are the standard normal density and distribution functions, respectively. For computation convenience, we solve the MLE problem by instead minimizing the negative log-transformation of L.