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

Sam Boysel

University of Southern California
Department of Economics

September 13, 2022

Outline

Introduction

- Open Source Software: Widely used, produced from disproportionate efforts of maintainers (Eghbal, 2020)
- Private provision of public goods can be sustained through benefit and/or cost heterogeneity (Bergstrom et al., 1986; Andreoni, 1990)
- Idea: peer contribution may influence benefits and/or costs of subsequent contribution
- **Questions:** Can peer effects effectively subsidize private provision of public goods? What is the value added by peer effects to private public good provision? Can peer influence sustainably distribute the cost of OSS maintenance?

Preview of Results

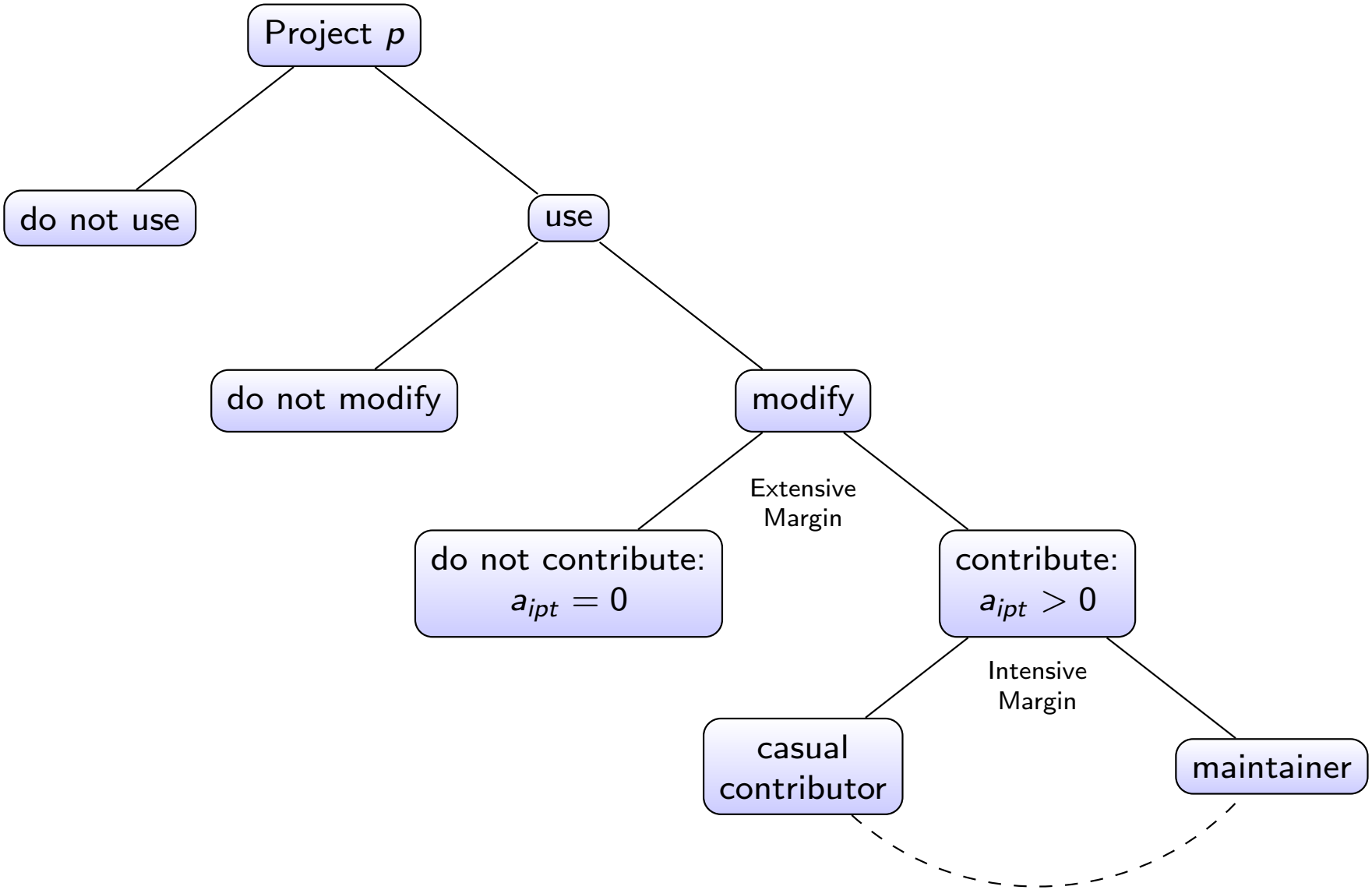
1. No strong evidence for contemporaneous, intensive margin peer effects
2. No strong evidence for peer influence in productivity
3. Heterogeneity across projects and time
   - Peer effects stronger in early days of GitHub, in smaller projects
4. Free-riding remains prevalent
   - Dominant contributors "crowd out" contribution from outsiders
5. Peer effects much stronger along extensive margin than intensive margin
6. Counterfactual: Extensive margin peer effects account for 56% of aggregate contribution over sample period

Setting

- Open Source Software is developed incrementally and collaboratively in public.
- Time-constrained developers contribute labor to write code.
- Socio-technical elements of the project community may influence contribution decisions
  1. Large-share contributors $\Rightarrow$ incentives to free-ride
  2. Some forces (critical segments of code, documentation, automation features, social interactions) may actually induce additional contribution.
- How can the net effect of these forces combine to sustain OSS projects?

▸ setting details

Setting: Decision Tree for an OSS user i at time t



Literature

- Why contribute to OSS? (Lerner and Tirole, 2002; Lakhani and Wolf, 2003; Eghbal, 2020)
- Theory on OSS project growth (Johnson, 2002; Athey and Ellison, 2014)
- Peer effects: productivity (Mas and Moretti, 2009), innovation (Fershtman and Gandal, 2011), open production (Zhang and Zhu, 2011; Slivko, 2014)
- Structural model: private public good contribution (Bergstrom et al., 1986) with selection (Heckman, 1979)

Data

Contribution patterns for sample of popular GitHub OSS projects. For agents $i \in \mathcal{N}$ (107,921) and OSS projects $p \in \mathcal{P}$ (2,287), we observe

- Contribution levels (number of commits) a_{ipt}
- Project quality (GitHub "stars") y_{pt}
- Time allocation ("GitHub active" days in month) x_{it}

for $t \in \{\text{April 2008}, \ldots, \text{June 2019}\}$

▸ descriptive statistics ▸ skewed contribution ▸ key unobservables

Reduced Form

Peer effects on contribution level (intensive margin)

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

δ is the marginal effect of peer contribution on individual contribution

- $a_{ipt} \geq 0$ is i's individual contribution
- $a_{-ipt} \geq 0$ is contribution of i's peers (sum)
- $\boldsymbol{X}_{ipt}$ includes fixed effects, cumulative sums, lags, and covariates like age, number of contributors, projects quality

Identification

Peer contribution likely suffers from *endogeneity bias* $\Rightarrow$ what can drive quasi-random variation in the contribution levels of my peers?

Agents are connected in a "project-mediated" network: we are peers if we contribute to the same project.

Peers-of-peers strategy: agent i's peers $j \neq i$ might change their contribution levels to p based on influences occurring in their outside options (e.g. projects $q \neq p$).

Key assumption: Agent i's peers-of-peers $k \neq i, j$ influence a_{ipt} only through influencing a_{-ipt}

Microfounded: operates through substitution and complementarity (as opposed to Bramoullé et al. (2009))

▸ identification details ▸ instrument definition

Reduced Form Estimates

Contemporaneous, intensive margin peer effects on contribution

	OLS Individual Commits			IV 2SLS Individual Commits		
	(1)	(2)	(3)	(4)	(5)	(6)
Peer Commits	0.0078	0.0065	0.0035	-0.0035	0.0131	0.0102
	(0.0011)	(0.0018)	(0.0034)	(0.0015)	(0.0054)	(0.0251)
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	0.1801	0.2268	-0.0007	0.1801	0.2267
1st stage F stat				6,520	1,151	64.37

- Controls: three lags of individual and peer commits, cumulative individual and peers commits, project quality, quadratic terms for project age and peer group size, and dummies for project ownership, membership, and firm affiliation.
- Fixed effects: individual, project, and year-month.

▸ high level findings ▸ project heterogeneity ▸ temporal heterogeneity ▸ beyond contemporaneous ▸ crowding out

Structural Approach

Motivation: microfoundations, counterfactual for value-added by peer effects $\Rightarrow$ suppose peers had no influence on contribution.

Features:

- Agent's contribution allocation decision (perfect information)
- Extensive margin: seection mechanism (Heckman, 1979)
- Intensive margin: static utility maximization subject to contribution constraint (Bergstrom et al., 1986)
- Peer effects embedded into each margin
- Contribution constraint disentangles productivity

▸ high level findings ▸ timing ▸ extensive margin ▸ intensive margin ▸ equilibrium ▸ peer effects
▸ estimation details ▸ estimates ▸ counterfactual

Discussion

- Punchline: peer effects can drive extensive margin OSS contribution but we cannot conclude they improve productivity
- Most users continue to free-ride on the efforts of dedicated maintainers
- Consistent with theory (Athey and Ellison, 2014) and anecdotal evidence (Eghbal, 2020)
- OSS is *digital infrastructure*
  - simple contribution $\neq$ sustained maintenance
- So what? Future directions
  1. Minor details: importance of non-code contributions, project "technical capital"
  2. Bigger questions: valuation of maintenance work (both for dependent projects and end users)

▸ policy recommendations

Details

Setting

- Software forges: SourceForge, GitLab, Bitbucket, and GitHub.
- Provide a common platform for collaboration and development of public goods
- Platform serves as an intermediary for contributors to propose changes, share documentation, file bug reports.
- Contribution measured in *commits*, atomistic modifications to project's codebase
- Contribution history tracked in *version control system* (e.g. `git`)
- Forking: users can copy code and develop an alternative version
- Deprecation: technology can often rapidly become outdated
- Example: https://github.com/JuliaData/DataFrames.jl

back

Data

Table: Descriptive Statistics

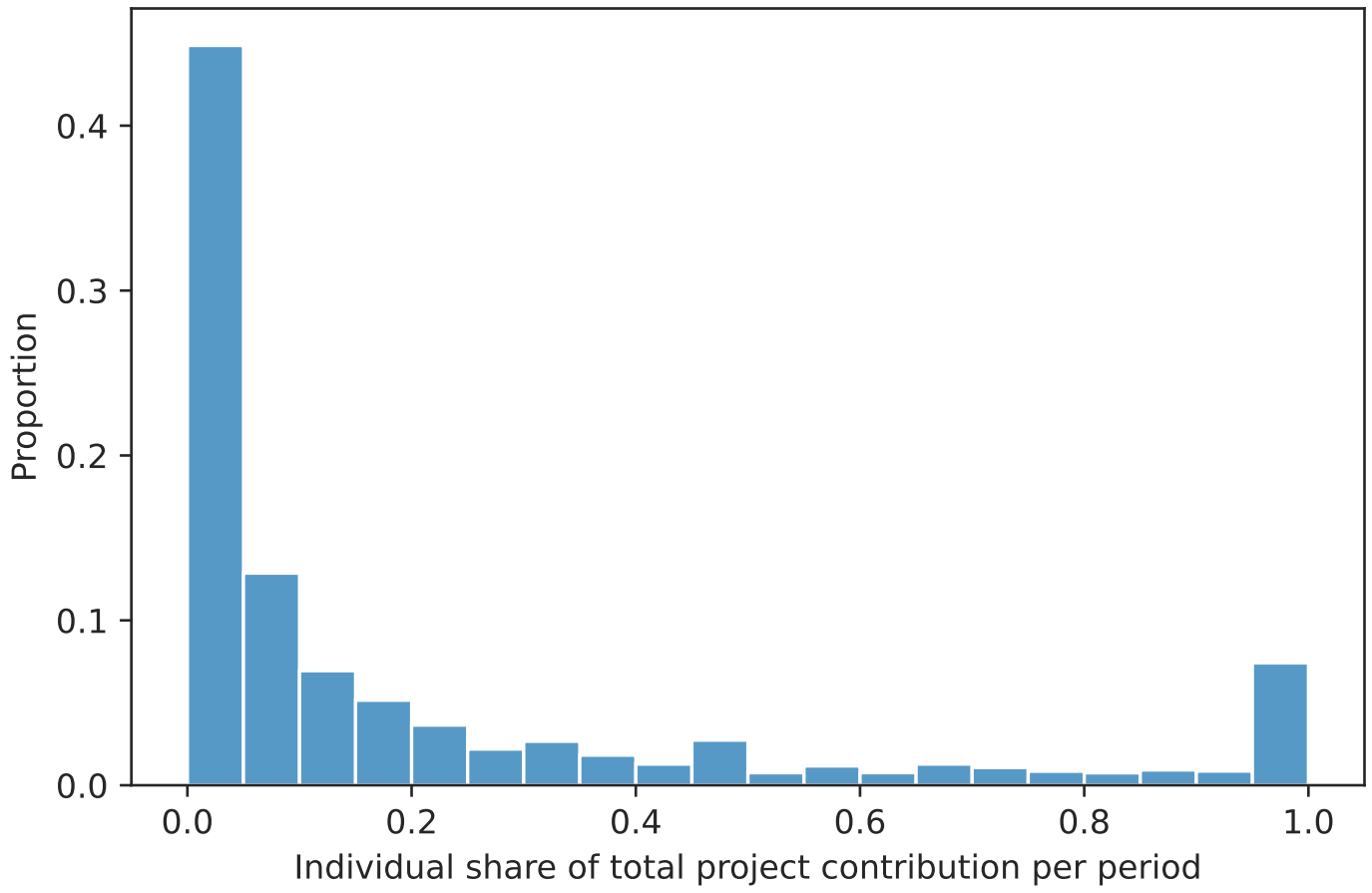
Measure	Notation	Obs	Mean	SD	Min	Median	Max
Individual commits	a_{ipt}	440,111	13	126	1	3	73,145
Peer commits	$a_{\text{-}ipt}$	440,111	188	398	0	59	73,160
Cumulative individual commits	$\tilde{a}_{ipt}$	440,111	256	1,076	1	23	186,447
Cumulative peer commits	$\tilde{a}_{\text{-}ipt}$	440,111	2,096	6,630	0	262	124,932
Number of peers	n_{ipt}	440,111	17	29	0	7	310
Cumulative GitHub Stars	y_{pt}	96,294	910	2,924	0	161	81,817
GitHub active days	gad_{it}	411,427	3.88	4.67	1	2	31

- Number of contributors: 107,921
- Number of OSS projects: 2,287
- Observed at monthly frequency from April 2008 through June 2019
- Takeaway: highly right-skewed contribution behavior

back

Data: Uneven Contribution Burdens

Figure: Distribution of Project-level Contribution Shares



back

Data: Key Unobservables

Note: (imperfect) proxies may exist for some of these features.

- Individual contributor characteristics
- Project-contributor match quality
- Private contribution benefits
- Aggregate project uptake and use value

back

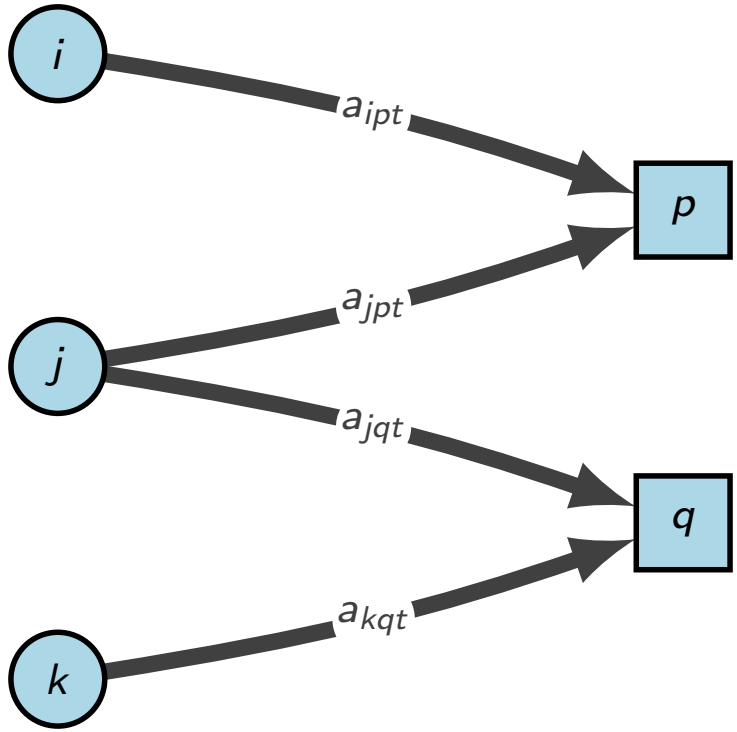
Identification Assumptions

1. **Independence:** peer-of-peer contribution levels fluctuate independently of *potential* individual and peer contribution levels
   - Agent i does not contribute to peers-of-peers projects
   - No coordination of contribution across projects
2. **Relevance:** $\text{Cov}[z_{ipt}, a_{-ipt} \mid \boldsymbol{X}_{ipt}] \neq 0$
   - conditional on other observables, peer contribution levels have some influence an individual's contribution level
3. **Exclusion:** $\text{Cov}[z_{ipt}, \epsilon_{ipt} \mid \boldsymbol{X}_{ipt}] = 0$
   - peers-of-peers contribution influences individual contribution *only* through peer contribution
   - No isolated contributors
   - Agents contribute to subset of projects but not all
4. **Monotonicity:** all agents find peer contribution either a substitute or compliment
   - Weaker condition: assume monotonicity holds within projects

back

Identification: Illustration I

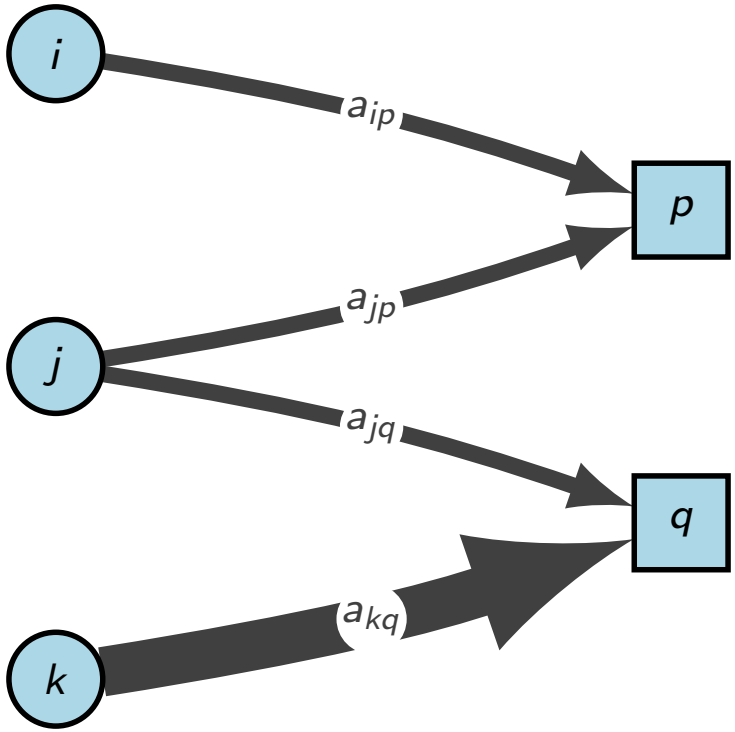
Figure: Peers of Peers



back

Identification: Illustration II

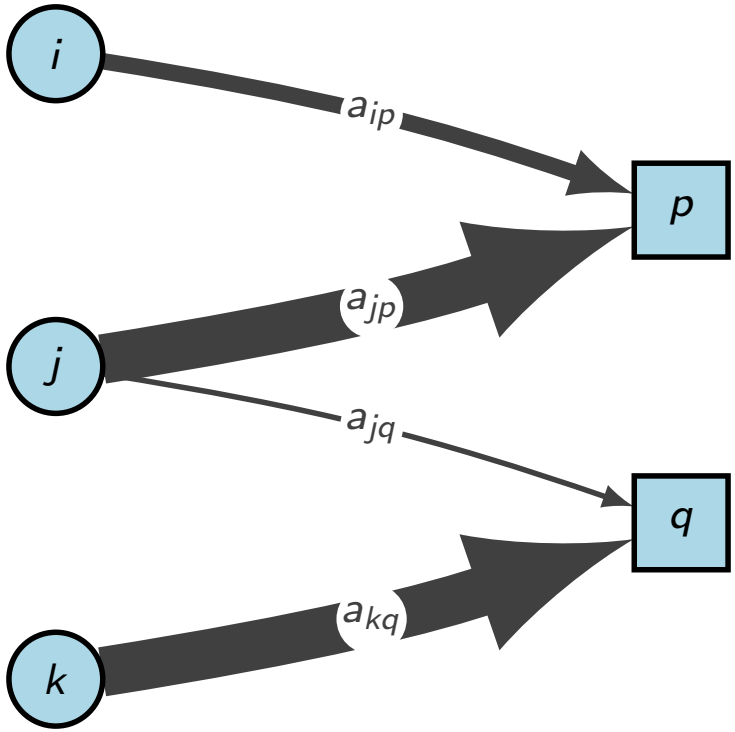
Figure: Suppose j's peer k increase contribution to project q



back

Identification: Illustration III

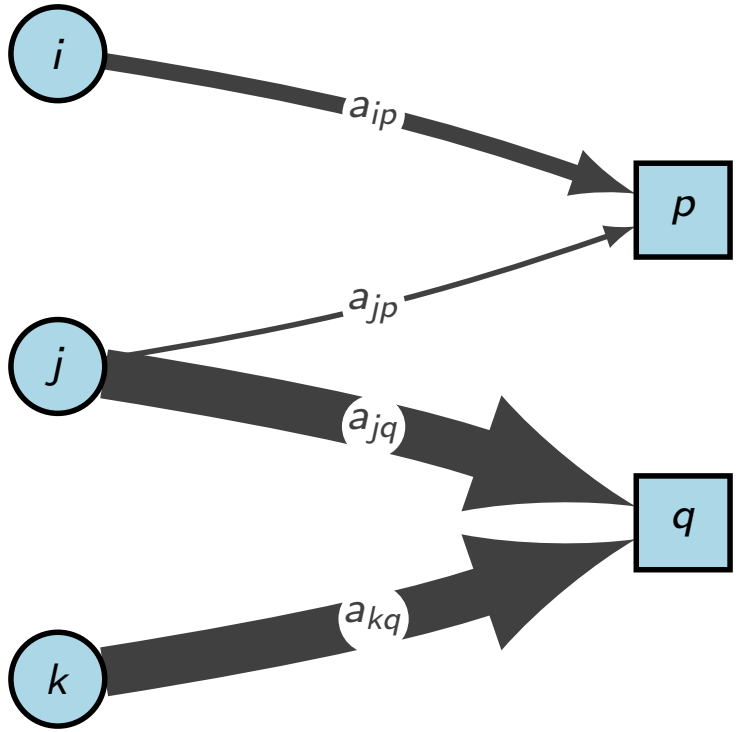
Figure: Case 1: j and k's effort in q are *substitutes*



back

Identification: Illustration IV

Figure: Case 2: j and k's effort in q are *complements*



back

Definition of Peers-of-peers Instrument

Denote the "peers-of-peers" instrument for peer contributions a_{-ipt} as z_{ipt}:

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

Hence z_{ipt} represents "aggregate contribution to projects shared by the peers of i's peers in project p in month $t-1$

back

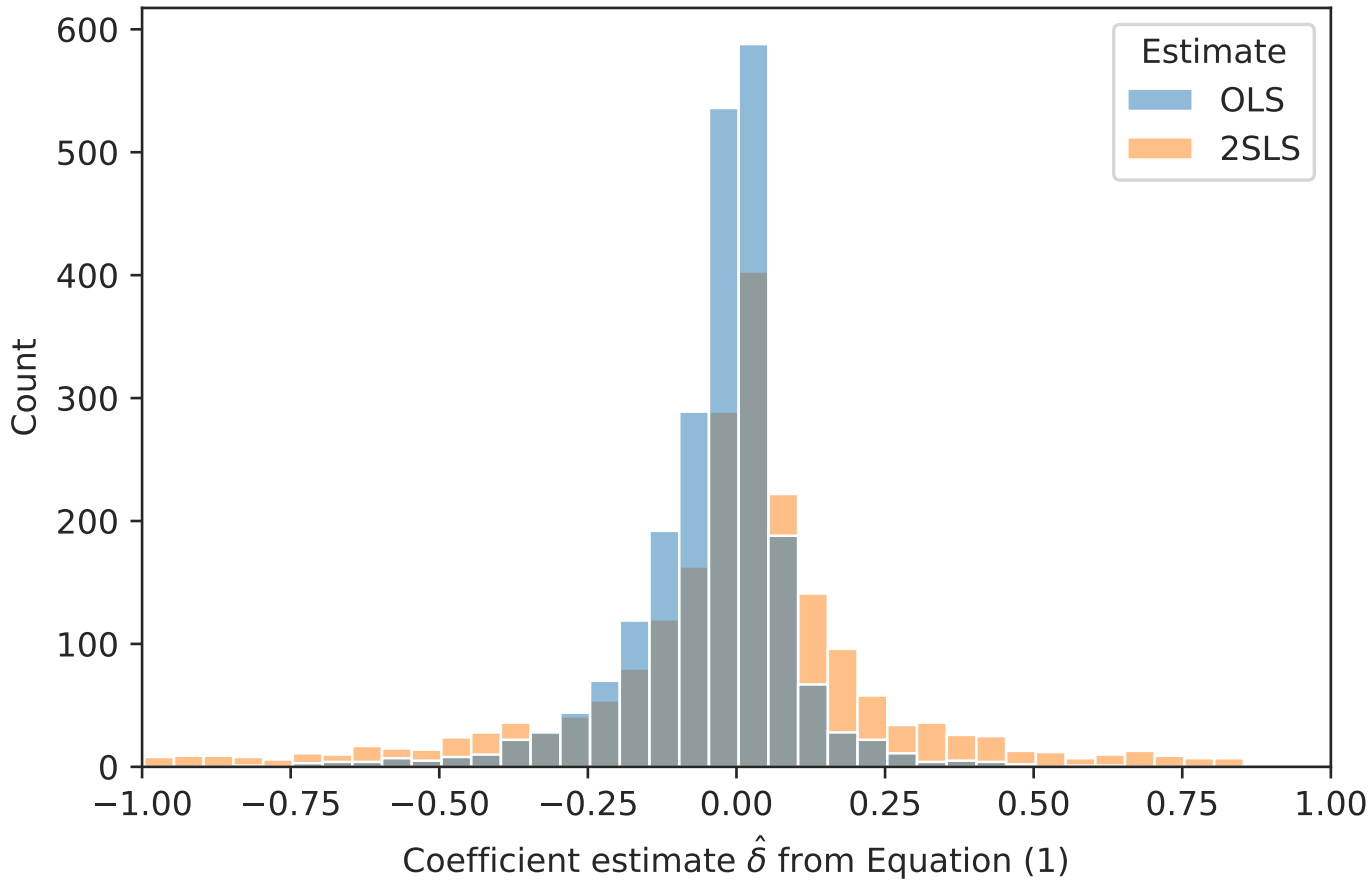
Reduced Form: High Level Findings

1. Little evidence for strong contemporaneous peer effects along the intensive margin, on average
2. Relatively symmetric distribution of intensive margin peer effects at the project-level about 0 (mix of positive and negative)
3. Peer effects stronger earlier in the sample period
4. Peer effects stronger when contemporaneousness assumption relaxed
5. Covariate Interactions: negligible influence
6. Cumulative and lagged individual contribution strong predictor of future contribution
7. Number of contributors strong predictor of participation (but at *smaller contribution levels*)
8. Project insiders "crowd out" outside contributions

back

Reduced Form: Project Heterogeneity

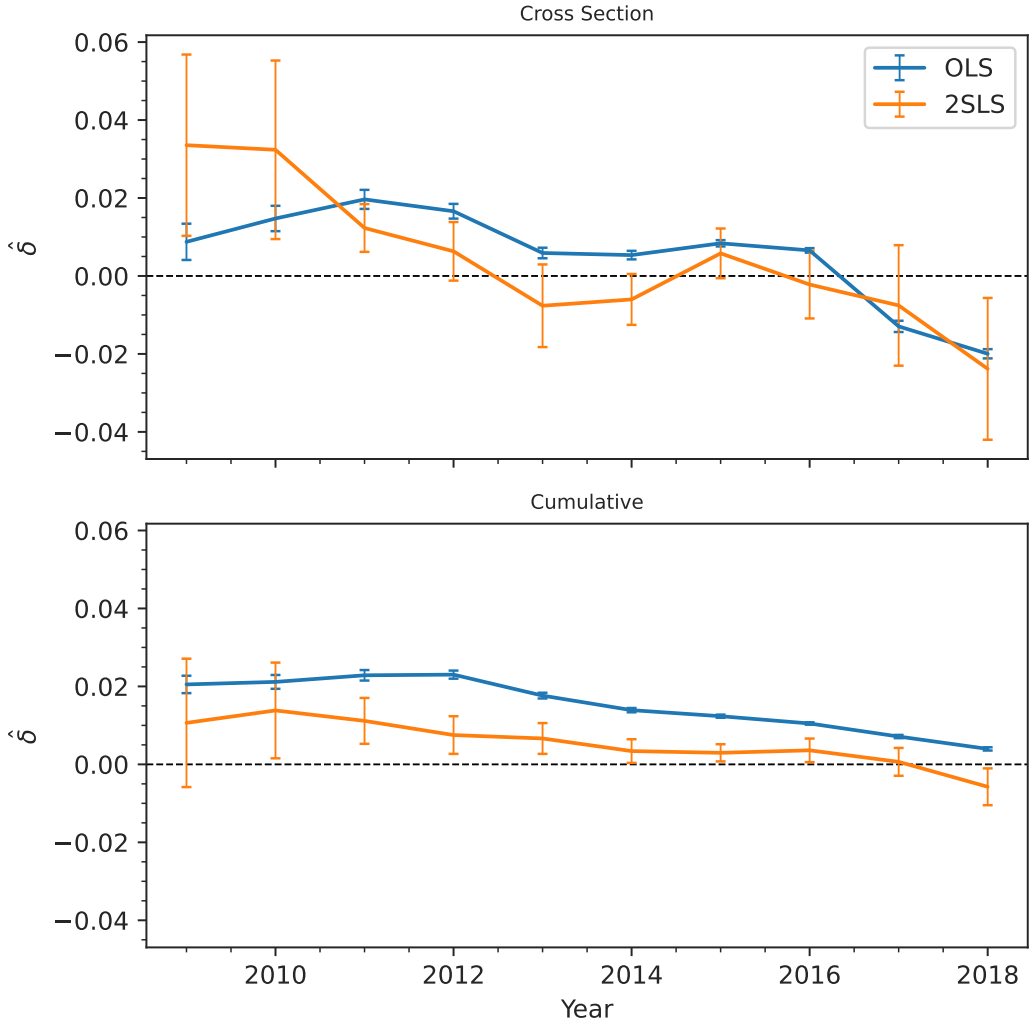
Figure: Estimates of contribution peer effects $\hat{\delta}$ at project-level



back

Reduced Form: Temporal Heterogeneity

Figure: Estimates of contribution peer effects $\hat{\delta}$ over sample period. Top: annual cross-sectional sub-samples. Bottom: cumulative sub-samples. back



Reduced Form: Beyond Contemporaneous Effects

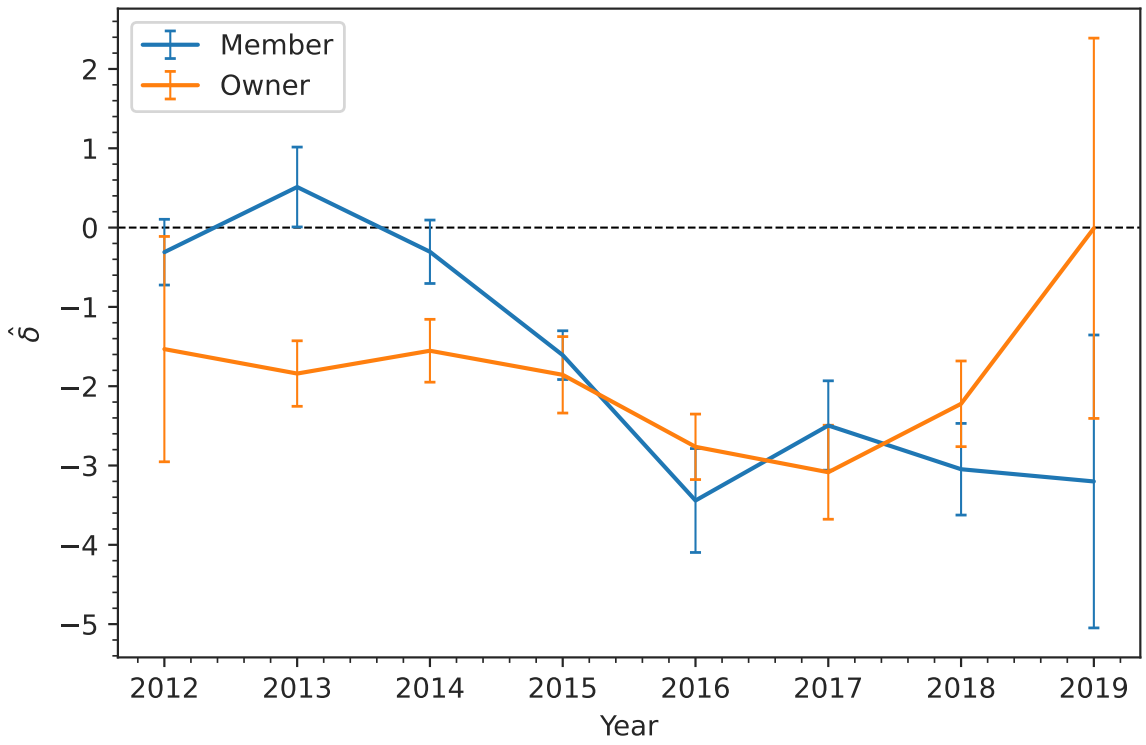
	OLS Individual Commits			IV 2SLS Individual Commits		
	(1)	(2)	(3)	(4)	(5)	(6)
Peer Commits	0.0139	0.0212	0.0068	-0.0145	0.0575	0.0881
	(0.0009)	(0.0042)	(0.0070)	(0.0018)	(0.0110)	(0.0914)
Individual Commits (cumulative)	-	-0.0051	-0.0063	-	-0.0074	-0.077
	-	(0.0020)	(0.0047)	-	(0.0021)	(0.0056)
Individual 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 $\hat{\delta}$ for the baseline specification in which aggregate peer commits from the preceding 3 months are regressed on individual commits for the subsequent 3 months.

back

Reduced Form: Crowding Out

Figure: Outsider (non-member or non-owner) contribution regressed on insider (member or owner) contribution at the project level



back

Structural Approach: High Level Findings

1. Extensive margin influences differ in direction
   - Positive effect: number of contributors in project
   - Negative effects: lagged and cumulative peer contribution
   - All effects decreasing in project size
2. Intensive margin peer effects (productivity and private benefits) small and centered around 0
3. Little evidence for peer influence over productivity
   - Strong negative correlation between benefit and productivity peer effects across projects
   - When agents contribute more, they do so at greater marginal cost
   - Pro-social forces trump productivity peer effects
4. Early sample periods: productivity and private benefits positively correlated $\Rightarrow$ greater net private benefits of contribution

back

Structural Approach: Timing

For each period t and project t, agent i determines optimal contribution as follows:

1. **Extensive Margin:** Learns a private fixed net benefit shock. Contribute to p if above threshold ($\Rightarrow a_{ipt} > 0$)
2. **Intensive Margin:** If $a_{ipt} > 0$, learns marginal private cost (i.e. productivity) and benefit shocks. Chooses optimal contribution level conditional on preferences and time constraint.

back

Structural Approach: Extensive Margin Decision

Agent i contributes to p at time t if (latent) fixed private benefits of contribution z_{ipt} exceed a threshold z_p:

$$a^{\star}_{ipt} > 0 \iff z_{ipt} \geq z_p$$

where z_{ipt} is a function of observables $\tilde{\boldsymbol{W}}_{ipt}$ and the shock $\epsilon^{z}_{ipt} \sim N(0,1)$:

$$z_{ipt} = \tilde{\gamma}' \tilde{\boldsymbol{W}}_{ipt} + \epsilon^{z}_{ipt}$$

Hence

$$\Pr(a^{\star}_{ipt} > 0) = \Pr(z_{ipt} \geq z_p) = \Phi(\tilde{\gamma}' \tilde{\boldsymbol{W}}_{ipt})$$

Note: the notation $\tilde{\boldsymbol{W}}$ is used as these observables will be later partitioned into peer and non-peer influences

back

Structural Approach: Intensive Margin Decision I

Project Quality

$$y_{pt} = b_{pt} \sum_{j} \sum_{s \leq t} a_{jps}$$

Contribution Constraint

$$x_{it} + \sum_{p \in \mathcal{P}} c_{ipt} a_{ipt} \leq 1$$

Preferences

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

back

Structural Approach: Intensive Margin Decision II

If agent i decides to contribute, she learns shocks for the marginal private benefit and costs of contribution: (v_{ipt}, c_{ipt})

Conditional on the shocks (v_{ipt}, c_{ipt}), i chooses a contribution level to solve the following utility maximization problem:

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

back

Structural Approach: Equilibrium

For each i, p, t:

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

and

$$\Pr(a^{\star}_{ipt} > 0) = \Phi(\tilde{\gamma}' \tilde{\boldsymbol{W}}_{ipt})$$

Therefore

$$E[a^{\star}_{ipt}] = \Phi(\tilde{\gamma}' \tilde{\boldsymbol{W}}_{ipt})(b_{pt} + v_{ipt} - c_{ipt})$$

back

Structural Approach: Peer Effects

Extensive Margin: γ

$$\begin{aligned} z_{ipt} &= \tilde{\gamma}' \tilde{\boldsymbol{W}}_{ipt} + \epsilon^z_{ipt} \\ &= \gamma' \boldsymbol{W}_{ipt} + \beta_z' \boldsymbol{X}_{ipt} + \epsilon^z_{ipt} \end{aligned}$$

where $\boldsymbol{W}_{ipt}$ includes (1) number of peers, (2) cumulative peer contribution

Intensive Margin: δ_v, δ_c

$$c_{ipt} = \delta_c \overline{c}_{\text{-}ipt} + \beta_c' \boldsymbol{X}_{ipt} + \epsilon^c_{ipt}$$

$$v_{ipt} = \delta_v \overline{v}_{\text{-}ipt} + \beta_v' \boldsymbol{X}_{ipt} + \epsilon^v_{ipt}$$

hence δ_v and δ_c captures correlation with peer benefit and productivity shocks.

back

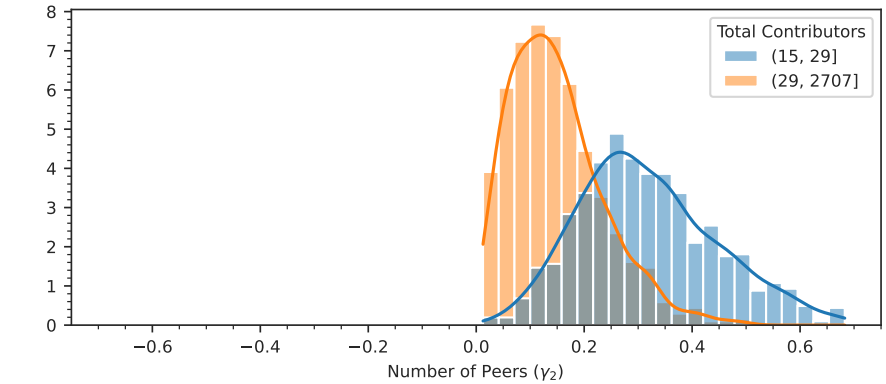
Structural Approach: Estimation Details

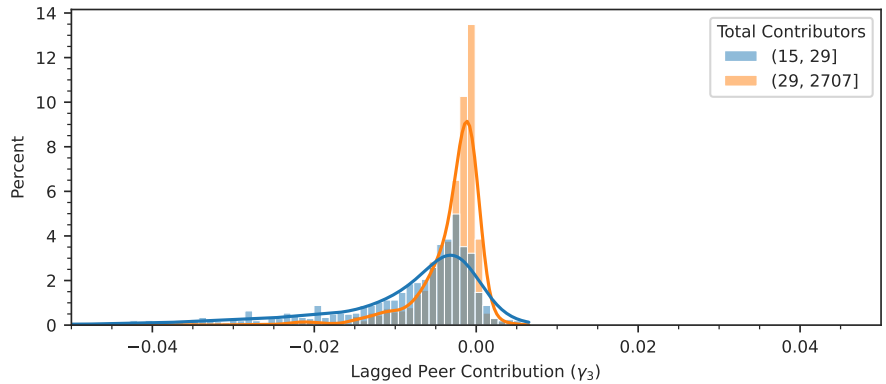
1. Assume disturbances are jointly normally distributed $(\epsilon^z_{ipt}, \epsilon^v_{ipt}, \epsilon^c_{ipt}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, independent and identically distributed between agents and time. Within the variance-covariance matrix $\boldsymbol{\Sigma}$, assume that $\sigma^2_z = 1$.
2. Given data (a_{ipt}, y_{pt}), recover $\boldsymbol{b} = (b_{pt})$ using project quality expression.
3. Given data $(a_{ipt}, y_{pt}, x_{it})$ and b_{pt}, recover shocks (v_{ipt}, c_{ipt}) using moment conditions (1) equilibrium contribution, (2) contribution constraint, and (3) project quality by means of GMM
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 MLE (Zhao et al., 2020)

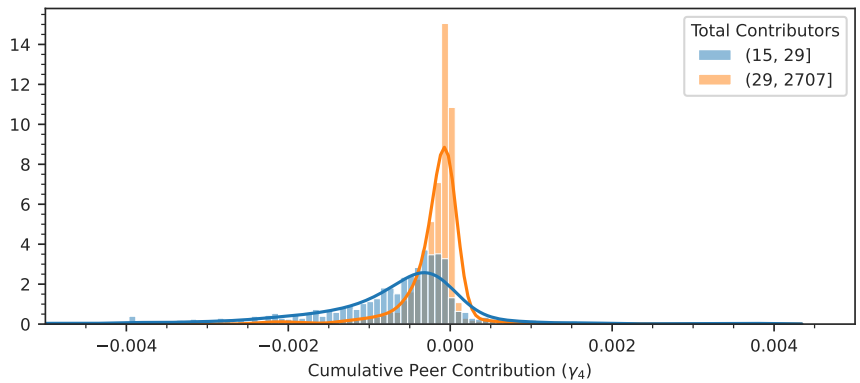
Parameters $\boldsymbol{\theta} = (\boldsymbol{b}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ completely characterize DGP

back

Structural Estimates: Extensive Margin Peer Effects

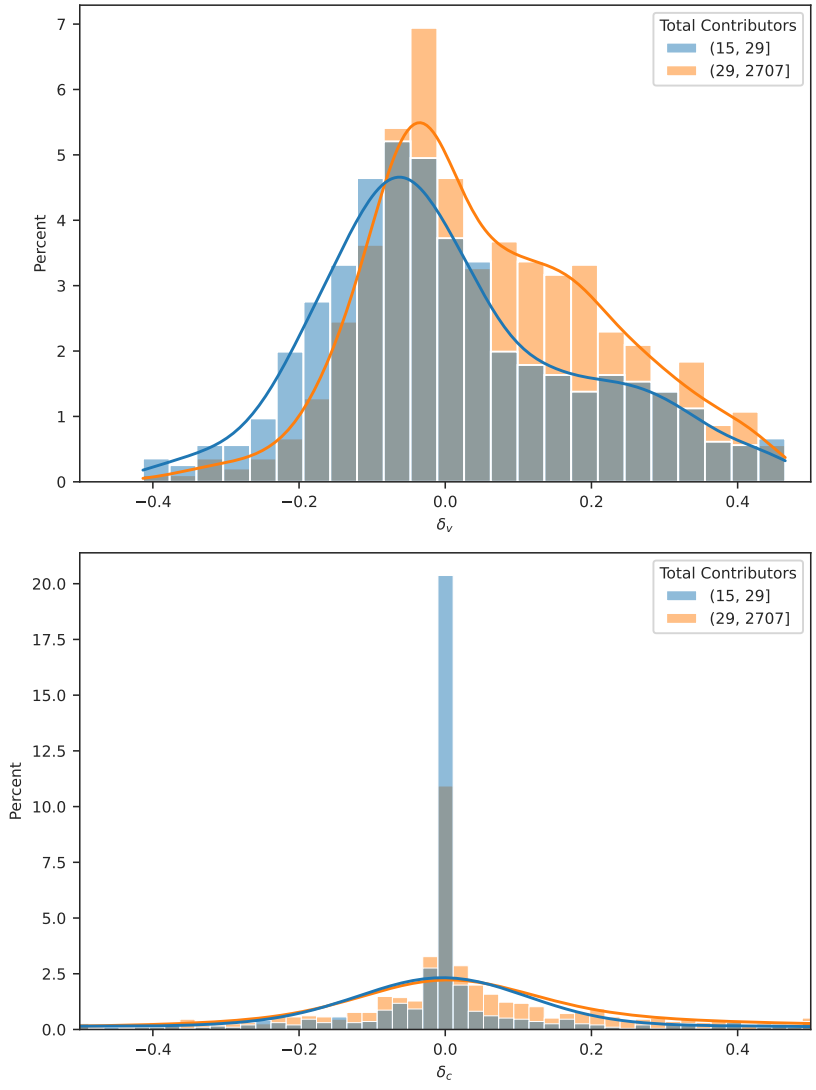






back

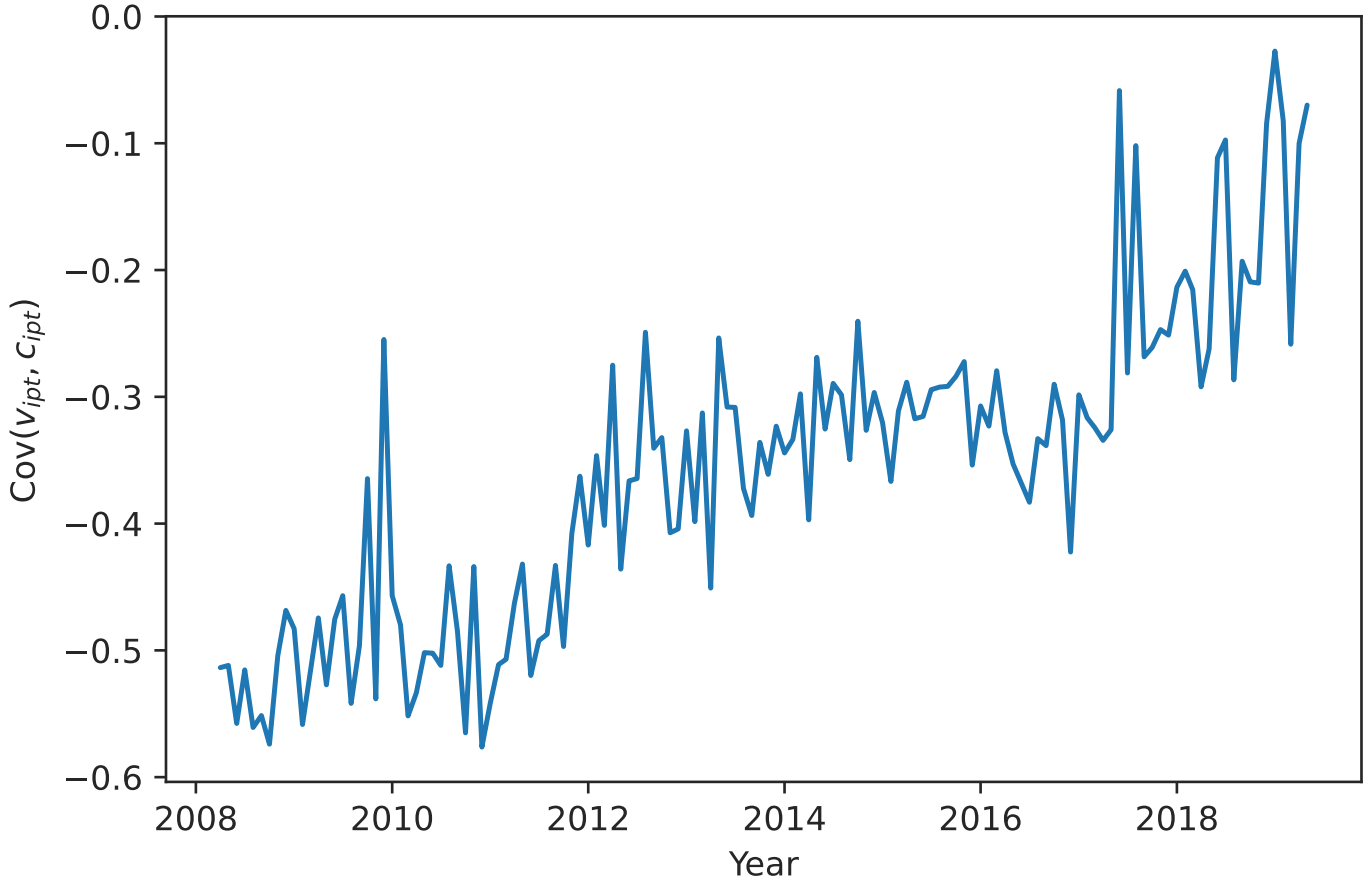
Structural Estimates: Intensive Margin Peer Effects



back

Structural Estimates: Benefit and Productivity Correlation

Takeaway: greater correlation between productivity and benefits in early days



◂ back

Structural Counterfactual: Value-added by Peer Effects

Suppose peers have no influence on individual contribution choices $\Rightarrow$ what is the counterfactual level of contribution?

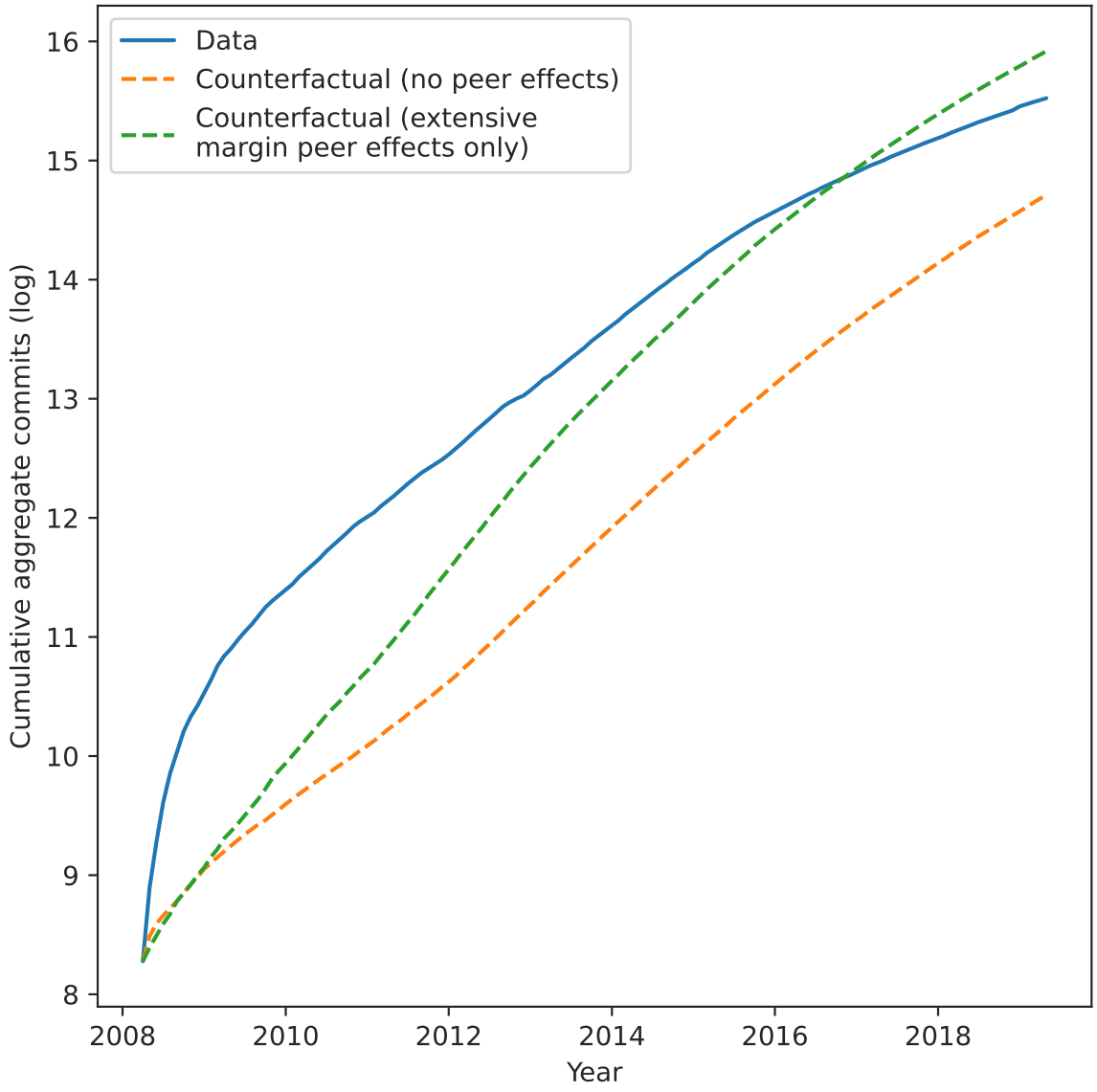
Operationalization: set estimated peer effects to 0 and re-simulate the DGP

Findings: Extensive margin peer effects explain roughly 56% of aggregate contribution (2.542 million hours or \$54-\$81 million USD for sample)

Intensive margin peer effects have negligible effects

back

Structural Counterfactual: Value-added by Peer Effects



back

Policy Recommendations

- Interest in cybersecurity at federal level (EO 14028)
- Silver bullet unlikely: solutions will come from private and public sector, both social and technical approaches
- Public support: fund maintenance work for widely depended upon OSS infrastructure (public goods rationale)
- Private innovation:
  - Business models to fund OSS: freemium, "open source code, paid support", private donations to sustain maintenance
  - Technical innovations: automation, continuous delivery, documentation and knowledge bases, online support communities

back