Discussion: The Private Value of Open-Source Innovation

Shan Huang (HKU)
shanhh@hku.hk

Agenda

- Research motivation & question
- Data & sample
- Methodology: Event-study valuation
- Key findings
- Robustness checks
- Discussion questions

Research Motivation & Question

- Open-source code is non-excludable—why do firms invest?
- Literature focuses on patents; limited large-scale evidence on OSI (open-source innovation) value.
- Research question:
  - Core question: What private shareholder value arises when a firm open-sources a repository?
  - Secondary: Which project, firm and market attributes amplify or dampen that reaction?

Correlational *not* Causal

- The research question is **novel, and empirically tractable**, making it a strong anchor for the study.
- **Question:** This paper builds more on correlations rather than causality.
- **Suggestion:** Can open source be the cause of high value of innovation?
  - A causal narrative can offer more actionable insights. For example, it can highlight value differences between similar innovations with and without open-source elements.
  - Demonstrating mechanisms through conducting more causal analyses can make a more powerful story.

Data & Sample

- 1,281 U.S. public firms (1997-2023) matched to 168k **GitHub repositories**
- 18 % of listed firms open-source yet represent 68 % of market cap & 80 % of R&D
- Repository ‘public’ timestamps identified from GitHub API events
- CRSP daily returns, patent & Compustat firm-level outcomes

Does GitHub Represent OSI Well?

- GitHub is now the world's dominant public code-hosting platform, so its logs are an **unusually rich lens on software-based innovation**—but the picture is **partial and sometimes distorted**.
- **Sample selection**:
  - People select certain projects onto GitHub as public repos (the paper findings also suggests)
  - Many corporate engineers push from personal accounts, leading to under-counting of corporate OSI.
- **Disclosure date ≠ invention date**.
  - Firms may spend years developing code privately, then open-source it only when strategic;
  - the "public" timestamp is therefore an imperfect innovation birthdate and can be endogenous.
- **High commit counts** may reflect refactoring rather than genuine novelty.

Methodology: Event-Study Valuation

- Abnormal return (CAR) over [0, +2] trading days around repository release
- Convert CAR to dollar value: CAR × firm market cap
- Aggregate across simultaneous releases at firm-day level
- Placebo dates & popularity validation bolster design

Robustness & Concerns

- Endogeneity: firms may open-source anticipating growth
  - Firms often *choose* the timing—e.g., align releases with developer conferences or earnings days—so the shock may not be exogenous.
  - IV suggested : the number of releases in the previous months or in the firm's close networks
- Mega-caps issue news constantly; their abnormal return signal can be swamped.
  - Excluding overlapping earnings announcements
  - Double-clustered SEs & FF-49 industry × year FE
- "Open-sourcing" often coincides with *product announcements.* A ±1-day window may still pick up those effects unless explicitly controlled.
  - **Placebo tests are good!**
  - **Multiple windows** – Report results for [-1,+1], [-2,+2], and [-5,+5]. True repo shocks should show a monotonic decay; confounded events often grow with the window.

Key Findings

- Mean private value ≈ US$0.84m per repository > Median 0.56 m; skewed distribution
- Aggregate private value 1997-2023 ≈ US$25 bn
- LLM-based complementarity scores negatively correlated, while novelty scores positively correlated with the value.
- Copyleft licenses, stand-alone projects, and low-competition industries drive higher values
- Repository value predicts 3-year growth in sales, employment, and patenting

Suggestions & Future Work

- These analyses are detailed and reasonably interpreted in the paper.
- I have following suggestions:
  - I would not describe them as determinants (causations) of value; rather, they illustrate heterogeneous effects (correlations).
  - For example, “Copyleft” is endogenous -- firms self-select, and license may signal project importance.
    - Instrumental variable: Use historical license preferences in an industry-peer network.
  - The GPT-based measure should be applied more carefully. A single zero-shot prompt to a large language model is fragile and opaque. You may apply more advanced techniques such as SFT (supervised fine tuning) to enhance the accuracy.

Discussion Questions

- Does the CAR capture stand-alone code value or broader strategic signaling?
- How large is selection bias in which projects go public on GitHub?
  - Can we model the selection process to debias the sample?
- Are copyleft licenses valuable due to strategic restriction or signaling?
- Could open-source be a leading indicator rather than a driver of firm growth?
- What policy implications arise for encouraging corporate OSI contributions?

Thank you!

shanhh@hku.hk