연구 데이터 공유

공유와 경쟁 사이

권석범
성균관대학교 기술경영전문대학원/시스템경영공학과

June 12, 2024

@수요 포럼

한국과학기술기획평가원

과학 진보를 위해 연구 데이터 공유는 중요할까요?

YES vs. NO

연구 데이터 공유의 중요성

연구 결과 재현과 검증

다양한 연구 수행

연구 부정 확인 & 예방

중복 연구 투자 예방

연구자 간 데이터 공유 활성화를 위한 다양한 정책/제도 고려

연구 데이터 아카이브 (e.g., University of Rochester)

학술지의 데이터 공유 정책

미국 NIH, NSF의 연구 자금 정책을 통한 데이터 공유 활성화

.... Etc.

연구 데이터 공유는 활성화되어 있을까요?

YES vs. **NO**

현실은...

“텅 빈 데이터 아카이브”

NEWS FEATURE DATA SHARING NATURE|Vol 461|10 September 2009

Empty archives

Most researchers agree that open access to data is the scientific ideal, so what is stopping it happening? **Bryn Nelson** investigates why many researchers choose not to share.

In 2003, the University of Rochester in New York launched a digital archive designed to preserve and share dissertations, preprints, working papers, photographs, music scores — just about any kind of digital data the university's investigators could produce. Six months of research and marketing had convinced the university that a publicly accessible online archive would be well received. At the time of the launch, the university librarians were worried that a flood of uploaded data might swamp the available storage space.

Six years later, the US$200,000 repository lies mostly empty.

Researchers had been very supportive of the archive idea, recalls Susan Gibbons, vice-provost and dean of the university's River Campus Libraries — especially as the alternative was to keep on scattering their data and dissertations across an ever-proliferating array of unintegrated computers and websites. "So we spent all this money, we spent all this time, we got the software up and running, and then we said, 'OK, here it is. We're ready. Give us your stuff'," she says. "And that's where we hit the wall." When the time came, scientists couldn't find their data, or didn't understand how to use the archive, or lamented that they just didn't have any more hours left in the day to spend on this business.

As Gibbons and anthropologist Nancy Fried Foster observed in their 2005 postmortem[1], "The phrase 'if you build it, they will come' does not yet apply to IRs [institutional repositories]."

A similar reality check has greeted other data-sharing efforts. Most researchers happily embrace the idea of sharing. It opens up observations to independent scrutiny, fosters new collaborations and encourages further discoveries in old data sets (see pages 168 and 171). But in practice those advantages often fail to outweigh researchers' concerns. What will keep work from being scooped, poached or misused? What rights will the scientists have to relinquish? Where will they get the hours and money to find and format everything?

> "We got the software up and running and said 'Give us your stuff'. That's when we hit the wall."
> — Susan Gibbons

Some communities have been quite open to sharing, and their repositories are bulging with data. Physicists, mathematicians and computer scientists use arXiv.org, operated by Cornell University in Ithaca, New York; the International Council for Science's World Data System holds data for fields such as geophysics and biodiversity; and molecular biologists use the Protein Data Bank, GenBank and dozens of other sites. The astronomy community has the International Virtual Observatory Alliance, geoscientists and environmental researchers have Germany's Publishing Network for Geoscientific & Environmental Data (PANGAEA), and the Dryad repository recently launched in North Carolina for ecology and evolution research.

But those discipline-specific successes are the exception rather than the rule in science. All too many observations lie isolated and forgotten on personal hard drives and CDs, trapped by technical, legal and cultural barriers — a problem that open-data advocates are only just beginning to solve.

One of those advocates is Mark Parsons at

ILLUSTRATIONS BY J.H. VAN DEN HOONDRICK

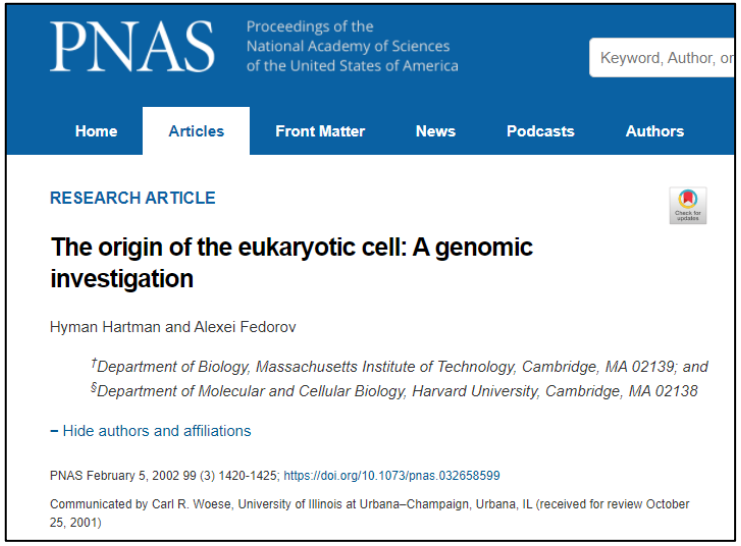
160

In 2003, the University of Rochester in New York launched a digital archive designed to preserve and share dissertations, preprints, working papers, photographs, music scores — just about any kind of digital data the university's investigators could produce. Six months of research and marketing had convinced the university that a publicly accessible online archive would be well received. At the time of the launch, the university librarians were worried that a flood of uploaded data might swamp the available storage space.

Six years later, the US$200,000 repository lies mostly empty.

왜?
세 명의 과학자 이야기

* Marshall, E. (2002). DNA sequencer protests being scooped with his own data. *Science*, *295*(5558), 1206-1207.


PNAS Proceedings of the National Academy of Sciences of the United States of America

Keyword, Author, or

Home | Articles | Front Matter | News | Podcasts | Authors

RESEARCH ARTICLE

The origin of the eukaryotic cell: A genomic investigation

Hyman Hartman and Alexei Fedorov

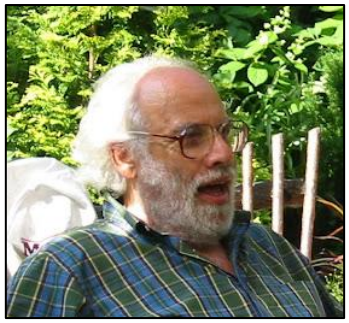
†Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; and §Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138

– Hide authors and affiliations

PNAS February 5, 2002 99 (3) 1420-1425; https://doi.org/10.1073/pnas.032658599

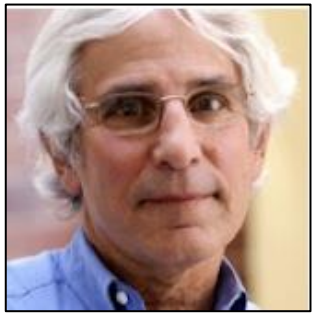
Communicated by Carl R. Woese, University of Illinois at Urbana–Champaign, Urbana, IL (received for review October 25, 2001)

Hartman, H., & Fedorov, A. (2002). The origin of the eukaryotic cell: a genomic investigation. *Proceedings of the National Academy of Sciences*, 99(3), 1420-1425



Dr. Hyman Hartman
@ MIT

Dr. Alexei Fedorov
@ Havard Univ



Dr. Mitch Sogin
@ Marine Biological Lab

"They used my DNA sequence data of *Giardia lamblia*. They never asked for my permission to use the data nor included me as a coauthor. **I am scooped**!

홈페이지 폐쇄
데이터 공유 중단

과학자들이 연구 데이터 공유를 꺼려 하는 이유?

과학자들이 기대하는 것과 걱정하는 것

- **기대하는 것**
  - ↑ 명성 & 연구자금 확보 가능성 &연구 협력 등
  - ↑ 연구 성과물의 가시성 (인용)
- **걱정하는 것**
  - 데이터 공유 비용 (Cost of storing, preparing, sharing data)
  - 같은 데이터를 활용하는 연구자와의 경쟁
  - Scooping 가능성과 보상의 불확실성

Q1. 연구 데이터 공유를 통해 (데이터를 공유하는) 과학자들은 이익을 얻고 있는가?

Q2. 연구 데이터 공유를 강제하는 정책은 (데이터를 공유하는) 과학자들의 연구 성과에 어떠한 영향을 미치는가?

연구 데이터 공유를 통해 (데이터를 공유하는) 과학자들은 (더 많은 인용을 통해) 이익을 얻고 있는가?

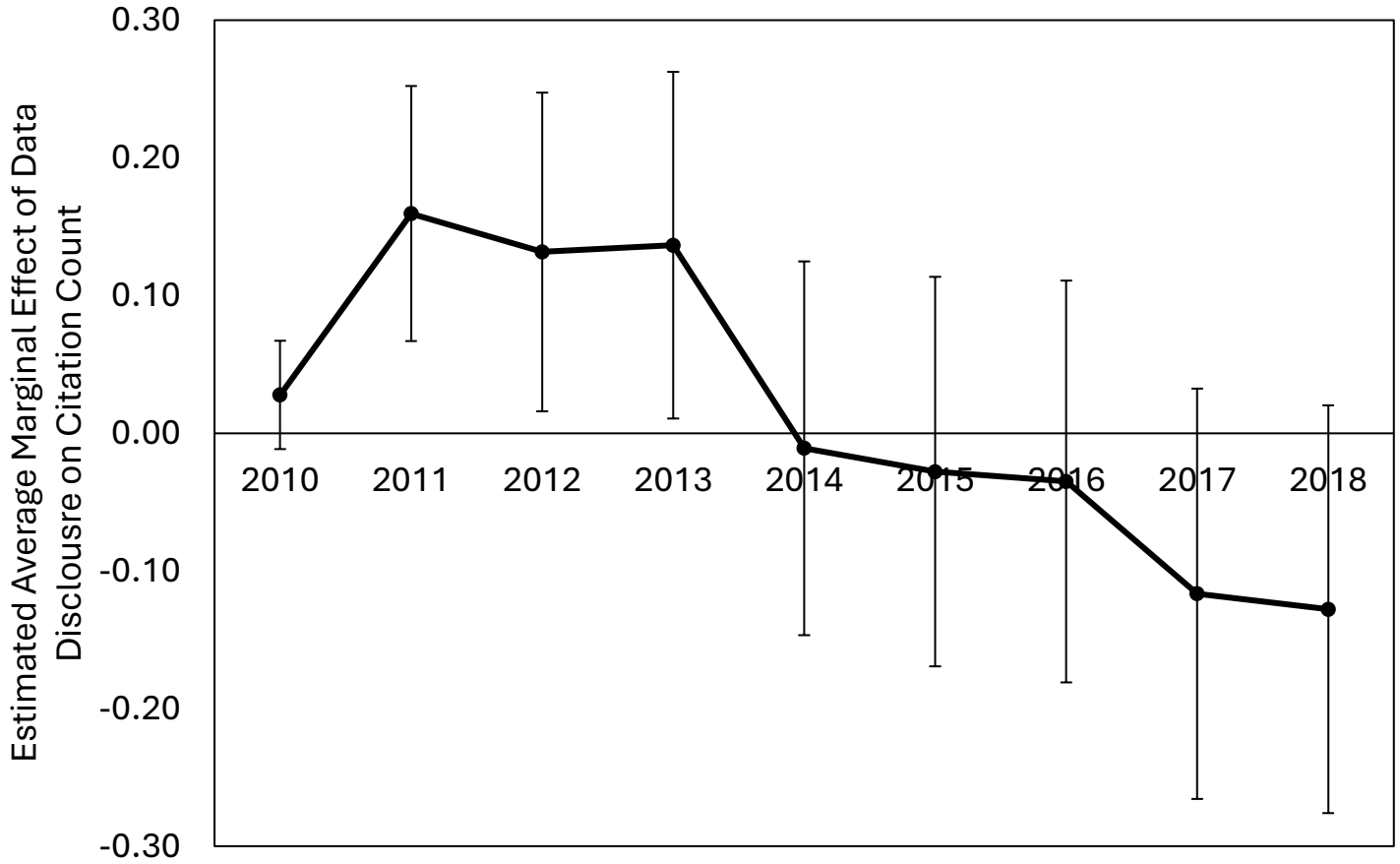
Kwon, S., & Motohashi, K. (2021). Incentive or disincentive for research data disclosure? A large-scale empirical analysis and implications for open science policy. International Journal of Information Management, 60, 102371.

데이터 공유 과학자의 연구 성과는 더 많이 인용됨

그러나 시간이 지날수록 더 적게 인용

15,271 data-disclosing papers+ 295,629 matched comparison group,

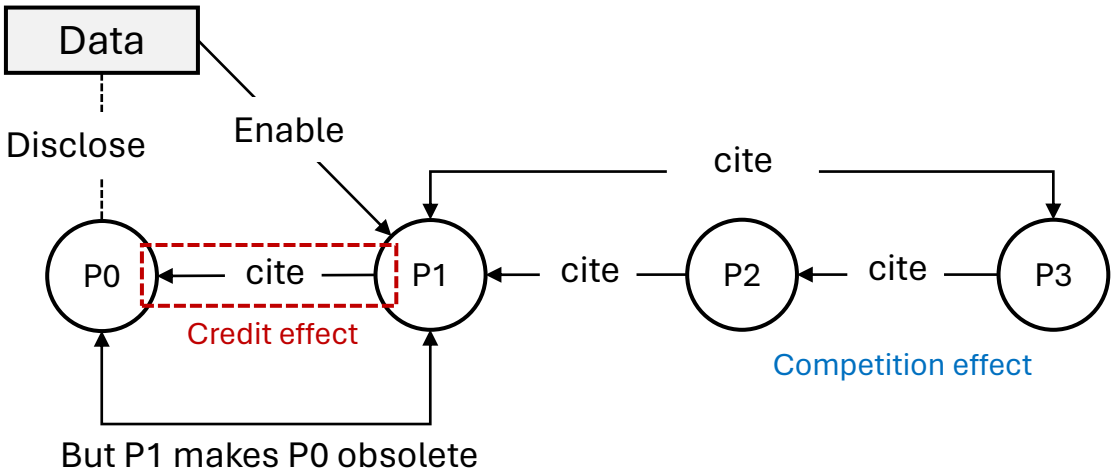
published in 1240 WoS-indexed journals in 2010

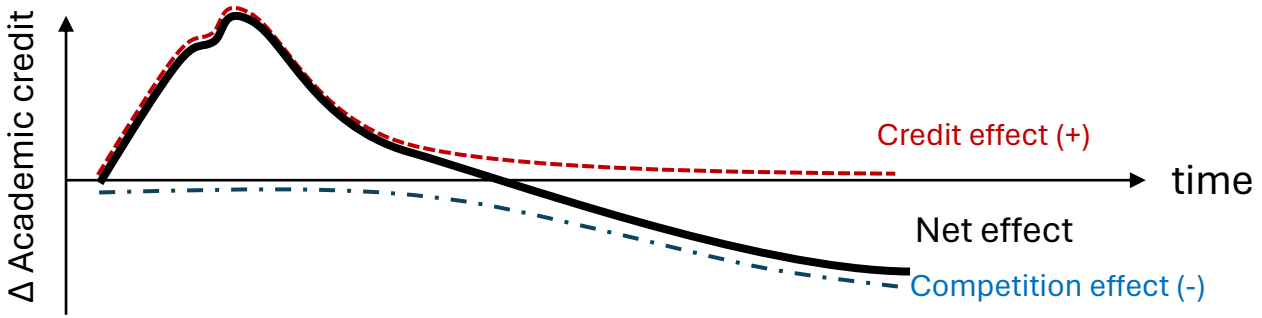


Error bar: 95% confidence interval computed with robust standard error

왜?

Credit vs. Competition Effect

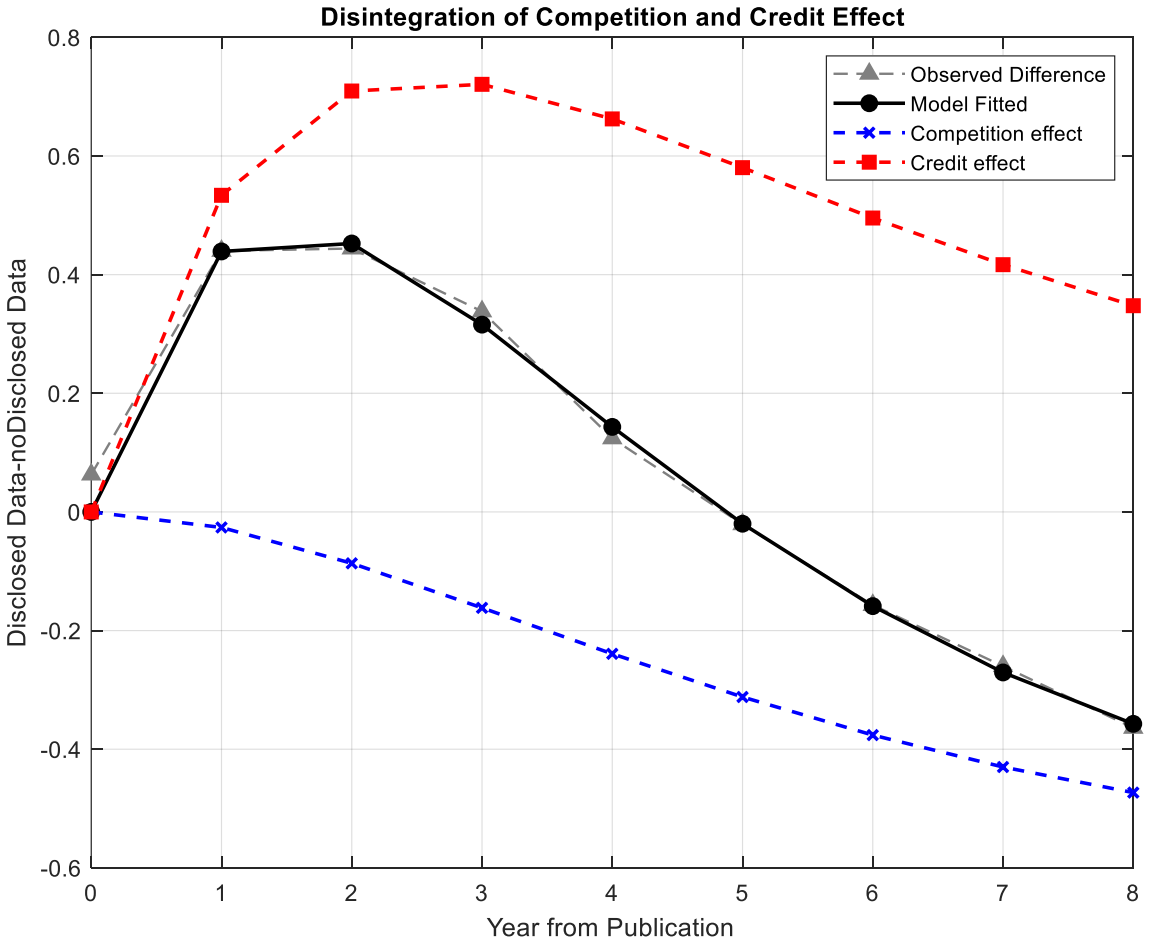




데이터를 공유한 후 초기에는 더 많은 인용 (Credit effect)
그러나, 시간이 지날수록 더욱 진보된 연구 결과 등장 (Competition effect)

추가 분석

Credit vs. Competition Effect Estimated



연구 데이터 공유를 강제하는 정책은 (데이터를 공유하는)
과학자들의 연구 성과에 부정적 영향을 미치는가?

Kwon, S. (2024). Competition or Diversion? Effect of Public Sharing of Data on Research Productivity of Data Provider. In *Academy of Management Proceedings* (Vol. 2024, No. 1, p. 12886). Briarcliff Manor, NY 10510: Academy of Management.

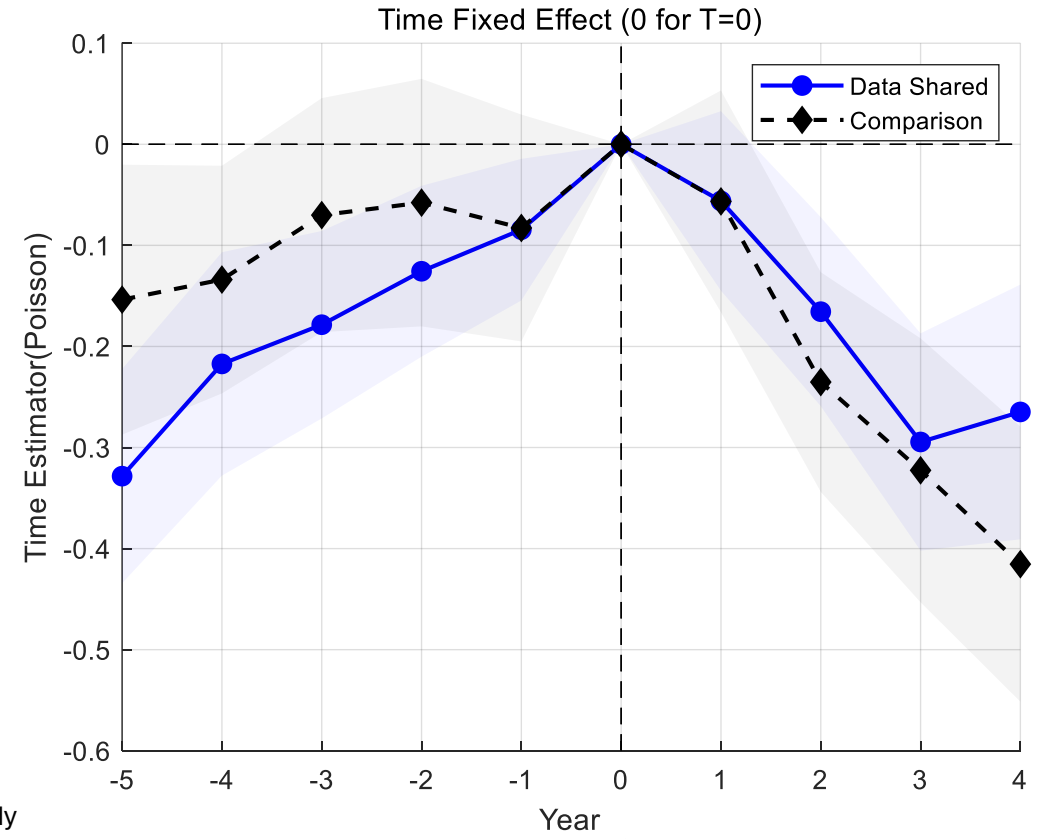
No Negative Impact

연구 성과는 저하 되지 않음

2007 and 2014 NIH Funding Policy Change (GWAS* + GDS** Policy)

1750 Data disclosing NIH projects + 1750 matched comparison group

Compare Publication performance since data disclosure



Erro rbar: 95% confidence interval computed with robust standard error

* Genome-Wide Association Study
** Genomic Data Sharing

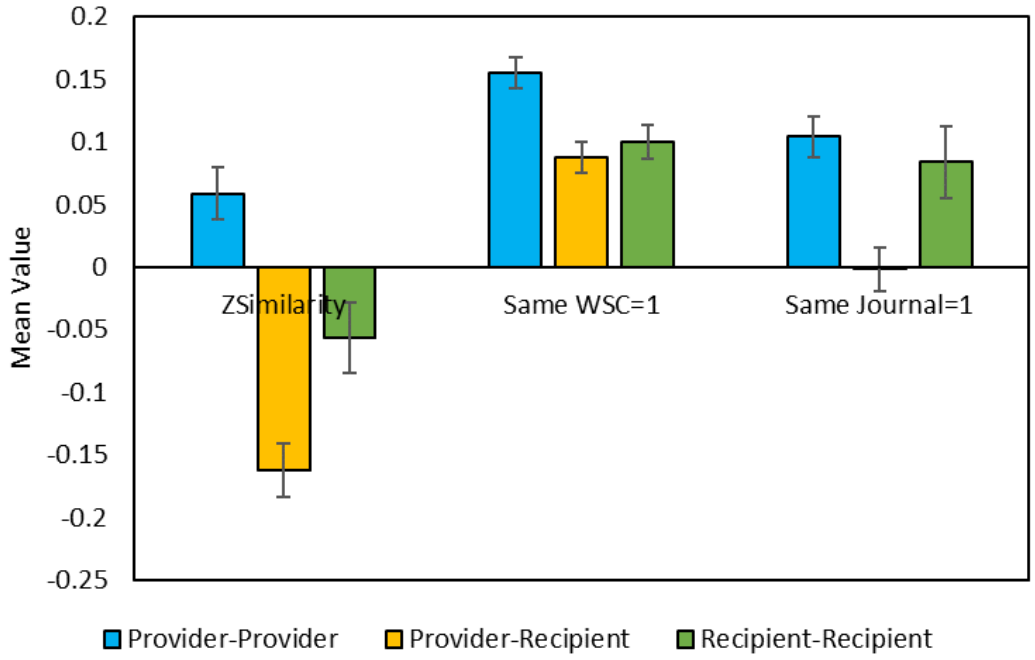
왜?

데이터를 활용하는 과학자들은, 데이터를 공개한 과학자와 "직접적인" 연구 경쟁을 하지 않음

Pair 1: provider-provider pairs

Pair 2: provider-recipient pairs

Pair 3: recipient-recipient pairs

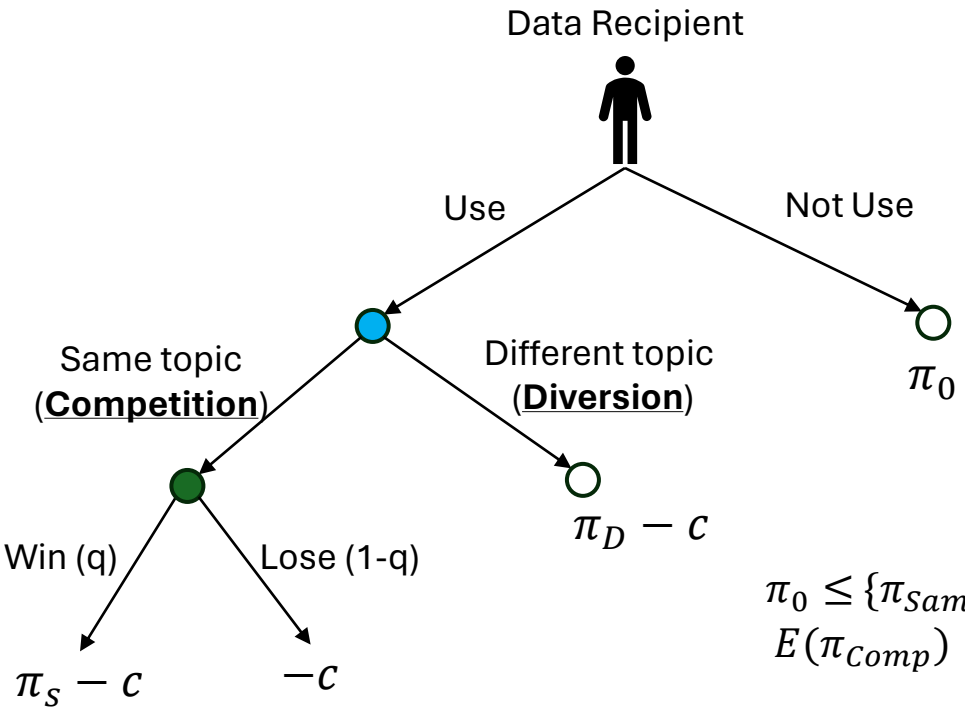


*wsc=web of science subject category

Error bar: 95% Confidence Interval computed using robust standard error
For I(Same Journal), 10 is multiplied due to the scale issue in the visualization

왜?

데이터를 공유하는 과학자와 직접적인 경쟁을 하는 것은 좋은 전략이 아님



$$\pi_0 \leq \{\pi_{SameTopic}, \pi_{DiffTopic}\}$$
$$E(\pi_{Comp}) = q\pi_s - \mathrm{c}, E(\pi_{Div}) = \pi_D - c$$

$$q \leq \frac{\pi_D}{\pi_S} = \tau \rightarrow \text{“Diversion”}$$
$$\mathrm{P}(Diversion) = f(q \leq \tau)$$

연구 데이터를 공개한 과학자와의 직접적인 경쟁 (Competition) 은 좋은 전략이 아닐 수 있음

→ **다른 연구 질문을 해결하는 것(“Diversion”)이 더 나은 전략**

결론 및 토의

I. 연구데이터에 대한 권리 강화?

- ✓ 연구 데이터에 대한 접근 권한 필요

  e.g., patent on invention, copyright on software

- ✓ "데이터 라이선싱" 제도? (Eisenberg, 2006)

II. 연구 데이터 공유 강제?

- ✓ 비활성화된 연구 데이터 공유에 대응한 "정책"의 필요성
- ✓ 연구 데이터 강제 정책의 부작용 가능성?

III. "경쟁"보다는 "다양한 연구" 수행을 유도할 수 있는 정책 필요

- ✓ 잘 설계된 제도는 "diversion"을 유도할 수 있음
  - (예) Data-embargo rule
  - 연구 데이터를 공유한 연구자의 현재 연구 프로젝트/성과물에 대한 정보
- ✓ 공개된 연구 데이터가 어떻게 "활용"되는지 더 많은 연구 필요
  - ✓ 데이터 활용자가 데이터 공유자와 "경쟁" 하는 때는?
  - ✓ 데이터 공유자가 진행중인 연구 내용에 대한 효과적인 inform 수단은?
  - ✓ 연구의 상업적 가치와 공유 데이터 활용 방향과의 관계는?