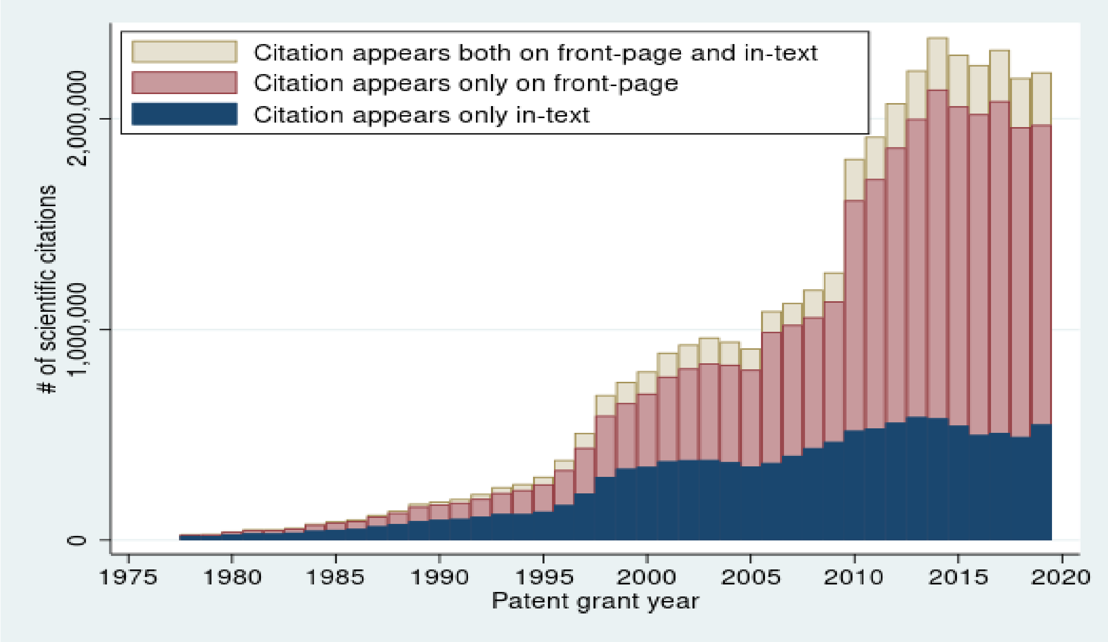
Linking Patent Citations to Scientific Articles to Understand the Heritage of Innovation

CAC

Project description

Scientific research plays a crucial role in driving innovation and economic growth, as it advances knowledge that leads to the development of new technologies, products, and services. One of the ways to measure the impact of scientific research on innovation is to track the number of times scientific articles are cited within the text of worldwide patents over extended periods of time. However, previous research was limited by computational constraints and methodologies, which only allowed for the counting of citations located on the front page of a patent. This prevented a comprehensive understanding of the "heritage of innovation" since in-text citations could not be adequately curated.

To address the limitations in previous studies, Matt Marx, a senior professor of personal enterprise at Cornell's Dyson School, and Aaron D. Fuegi, senior graphic analyst at Boston University, collaborated to curate and analyze a comprehensive dataset consisting of 16.8 million citations from the full text of U.S. patents since the first patent was issued in 1836, as well as European patents starting in 1978. They utilized a combination of hand-tuned heuristics to achieve the best precision and recall, and also incorporated fuzzy matching. This analysis was scaled up to cover the entire patent/article corpus. The resulting Version 1 of the dataset was published in the *Journal of Economics & Management Strategy* in 2022 under the title "Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations" and was generated using Boston University's shared computing cluster.



This chart illustrates the evolution of both front-page and in-text citations in European patents over time. Marx also replicated previous studies of U.S. patents using full-text citations and found that patents are 40% closer to the academic/industry interface. Additionally, he discovered that the percentage of NIH grants that gave rise to commercial applications was understood by approximately 20%.

A complete rebuild of Marx's data set, referred to as Version 2, was curated and scaled on Cornell University's Red Cloud. This new version became necessary when Microsoft's Academic Graph project ended. To rebuild the data set, Marx's team used Red Cloud and linked it to OpenAlex, a new and open-source comprehensive index of scholarly papers, citations, authors, and journals. OpenAlex was beta-launched in 2022 and today serves as a replacement for the original Microsoft Academic Graph.

CORNELL FACULTY

Matt Marx

Charles H. Dyson School of Applied Economics & Management

CAC SERVICES

- Virtual Cluster Running in Red Cloud
- Python Virtual Environment
- Slurm Scheduler
- Jupyter Notebooks
- 50TBs Storage

CAC services

Cornell's Center for Advanced Computing (CAC) provides application consulting services tailored to individual research needs and operates a scalable cloud system called Red Cloud. To facilitate Marx's research, CAC consultants met with his team to understand their computing and software requirements and deployed a virtual cluster called "Marx1" in Red Cloud.

Virtual clusters in Red Cloud are an efficient alternative to traditional high performance computing clusters as they eliminate much of the overhead involved in setting up an individual cluster and can scale in minutes, saving time and effort. Recognizing these advantages, Marx selected the proposed virtual cluster in the cloud solution, and CAC helped to port the Sun Grid Engine (SGE) queuing system to the Slurm job scheduler.

CAC staff assisted in software installation, including Java, Perl modules, Fuzzy-Match, and Python developer packages. They also helped to resolve issues with GROBID, a machine learning system for document extraction. Additionally, CAC tested a strategy for running Jupyter Notebooks on a head node and provided step-by-step instructions on how to do it. Marx stated, "We wanted to achieve our results faster and more efficiently, and CAC helped make that happen." Marx has used 1,745,000 Red Cloud computing hours and 50 terabytes of storage to date. An interactive queue was added to allow Marx to do post-processing.

Results

The patent-to-paper citations dataset Marx and his lab ported to the CAC virtual cluster and rebuilt using the OpenAlex replacement for the Microsoft Academic Graph has been deeply impactful on the scientific community. The dataset has been downloaded more than 65,000 times by researchers around the world. Dozens of articles have been published using Marx's dataset. Moreover, the IMF's 2022 World Economic Outlook forecast relied deeply on Marx's dataset for its analysis.

A follow-on project, done exclusively on the CAC virtual cluster, extends the patent-to-paper citations to identify which patents not only cite a paper but are equivalent in content. These "patent paper pairs" capture the commercialization of science. To find pairs, Marx and his lab found instances where the inventors on the patent overlapped with authors on the paper (i.e., a self-citation) and also where stretches of the paper's abstract were copied-and-pasted into the text of the patent. This self-plagiarism is an advance over traditional text-similarity measures.

ABOUT CAC

The Cornell University Center for Advanced Computing is located in Ithaca, NY. With a professional staff of computational and data scientists, cyberinfrastructure specialists, and a breadth of PhD-level expertise in astronomy, biology, computer science, information science, mathematics, physics, and more, CAC is creating tomorrow's software, data, and computing solutions today.

CAC INQUIRIES

Susan Mehringer
Associate Director, Consulting
Cornell University Center for Advanced Computing (CAC)
shm7@cornell.edu
www.cac.cornell.edu

APPLIED ECONOMICS & MANAGEMENT

Matt Marx
Bruce F. Failing, Sr. Professor of Personal Enterprise, Dyson School
Cornell University S.C. Johnson College of Business
mmarx@cornell.edu