DISCERN 2.0: Extending and Enhancing the DISCERN Dataset

Ashish Arora[*] Sharon Belenzon[†] Larisa Cioaca[‡] Lia Sheer[§]
Dror Shvadron[¶]

November 20, 2023

Abstract

This paper introduces DISCERN 2.0, an extensive update to the original DISCERN dataset. The update extends the temporal coverage up to 2021, incorporates improved firm and subsidiary data, and integrates scientific publications and patent data using the OpenAlex and PatentsView datasets. These improvements result in a more comprehensive and accurate coverage, evidenced by increases in the number of ultimate owner firms, subsidiaries, patents, and scientific publications. The preliminary results are set to be presented at the NBER Innovation Information Initiative (I3) meeting.

[*] Arora: Duke University, Fuqua School of Business and NBER
[†] Belenzon: Duke University, Fuqua School of Business and NBER
[‡] Cioaca: Duke University, Fuqua School of Business
[§] Sheer: Tel-Aviv University, Coller School of Management
[¶] Shvadron: Duke University, Fuqua School of Business

1 Intro

We are pleased to present a comprehensive extension and enhancement of the DISCERN dataset (Arora et al., 2021b,a). Since its introduction in 2020, DISCERN has become a valuable tool for studying corporate innovation and patenting activities.[1] This is evidenced by over 16,000 downloads[2] and its use in publications, including studies published in the *American Economic Review*, *Management Science*, and *Research Policy*, among other journals.[3]

For this update, we have enriched the dataset in several ways. We have extended the coverage to include additional years, widened the coverage of subsidiary data, improved the quality of matches to patents, and added open-access matches to scientific publications. These enhancements broaden the dataset's utility, offering researchers a more comprehensive, precise, and reliable data source for studying the U.S. corporate innovation landscape.

2 Methods

Our current update includes data improvements on multiple fronts:

1. ***Firm panel:***

   (a) ***Sample period:*** We have extended the dataset to include six additional years. DISCERN 2.0 now covers firms between 1980 and 2021.

   (b) ***Ultimate owner (UO) data:*** We have updated the data on U.S.-headquartered, publicly traded ultimate owners up to 2021 using Compustat data (Standard & Poor's, 2022) while incorporating name changes (Center for Research in Security Prices, 2022) and dynamic ownership changes (Securities Data Company Platinum, 2022).[4] DISCERN 2.0 now includes both firms that patent and firms

[1] The original DISCERN dataset extended and replaced its predecessor, the historical NBER patent dataset (Hall et al., 2001; Bessen, 2009).

[2] The original DISCERN dataset can be downloaded from `https://zenodo.org/record/4320782`.

[3] Examples of published research papers based on the DISCERN dataset: Arora et al. (2021a, 2023a,b); Sheer (2022); Bahar et al. (2023); Papageorgiadis et al. (2023); Kini et al. (2023); Arts et al. (2023).

[4] For example, in 2017 the two largest publicly traded chemical companies, Dow and Dupont, merged. The combined company was renamed DowDuPont. In 2019, less than two years after its formation, the company dissolved, and three new publicly traded spinoffs were formed: the materials science division Dow, the specialty products division DuPont, and the agriculture division Corteva. The dynamic nature of our data traces these companies from two separate pre-merger entities, to the merged entity DowDupont, to the three spinoff companies.

that do not patent, allowing users to select the analysis subsample that fits their research needs.

(c) ***Subsidiary data:*** We have transitioned to using Subsidiary Data from WRDS (Wharton Research Data Service, 2023) as our primary source of UO-subsidiary ownership linkages. This dataset contains ownership relationships for firms filing with the U.S. Securities and Exchange Commission (SEC) between 1995 and 2022.[5] To overcome the challenge of identifying subsidiaries over time, we used the GVKEY-CIK Link Table from WRDS to map a company's unique identifier in Compustat (GVKEY) to all the historical CIKs under which it submitted filings to the SEC.[6][7]

WRDS data offer two clear advantages over the Orbis ownership data (Bureau van Dijk, 2018) used in the original DISCERN. First, the data coverage begins in the mid-1990s, compared to the early 2000s for ownership links in Orbis. Second, WRDS relies on companies' official reports of significant subsidiaries, thus ensuring a higher degree of reliability.[8] Similar to the UO extension, DISCERN 2.0 now includes both subsidiaries that patent and subsidiaries that do not patent.

2. ***Patents:***

(a) We have transitioned to using the PatentsView dataset (U.S. Patent and Trademark Office, 2022) as our primary source of patent data. This dataset covers all patents granted by the United States Patent and Trademark Office (USPTO) and also includes pre-grant patent publications. PatentsView offers clear benefits over PatStat, which was used in the original DISCERN. It is free for bulk download and easily accessible through either a USPTO API or the Google

[5] These relationships are parsed from exhibits attached to a variety of filing types (10-K, 10-Q, etc.), but rely primarily on Exhibit 21, Subsidiaries of the Registrant, filed as part of Form 10-K. This exhibit lists all existing significant subsidiaries owned by the company—either directly or indirectly through another subsidiary—including (a) its name and (b) its jurisdiction of incorporation.

[6] A CIK is a unique 10-digit identifier that the SEC's computer system assigns to individuals and corporations who file disclosure documents with the SEC. When a company's legal status changes—or in other situations involving corporate restructuring, spinoffs, bankruptcies, or mergers and acquisitions—companies might start reporting under a different CIK number.

[7] SEC filers often change their legal company names and other identification information, such as CIK, CUSIP, and exchange ticker symbol. For example, COOPER INDUSTRIES INC, a manufacturer of electrical lighting and wiring equipment, filed annual reports with the SEC under both CIK 0000024454 (during 1993-2001) and CIK 0001141982 (during 2003-2011), before being acquired by EATON CORP in 2012.

[8] The names of particular subsidiaries may be omitted from Exhibit 21 if the unnamed subsidiaries, considered in the aggregate as a single subsidiary, would not constitute a "significant subsidiary" as of the end of the year covered by the 10-K report.

Patents Public Data platform, making it a convenient data source for researchers.

(b) We have made significant enhancements to the procedure for matching firm names to patent assignee names. This updated procedure now considers potential matches derived from two distinct algorithms. As with the original DISCERN, we employ the Term Frequency - Inverse Document Frequency (TF-IDF) algorithm for fuzzy string matching. TF-IDF prioritizes matches with unique words (e.g., "Google") over those with words frequently found across many firms (e.g., "Incorporated"). Furthermore, this update introduces the Nearest-Neighbor (NN) TF-IDF algorithm, which accommodates typos at the word level. For instance, standard TF-IDF wouldn't recognize "MICROSOFT" and "MICSOSOFT" as referring to the same firm. In comparison, the NN version of the algorithm can consider these strings as potential matches. By leveraging these algorithms, we generate fuzzy candidate matches for each assignee name and then manually review them to eliminate any false positives.

(c) We have refined our approach to more comprehensively track changes in patent ownership using the USPTO Patent Assignment Dataset (PAD). In the original DISCERN, we focused on monitoring patent ownership changes based on events at the firm level. In this update, we not only accounted for dynamic changes in firm-level ownership but also tracked the movement of individual patents through re-assignment between the firms in our sample. By integrating the PAD dataset, we ensure a more accurate representation of patents entering, exiting, or changing hands in our data.

3. ***Scientific publications:***

   (a) We have made a significant shift by adopting the OpenAlex dataset (Priem et al., 2022) as our source of data on scientific publications. When developing the original DISCERN, we relied on the Web of Science (WoS) dataset to gather data on scientific publications. However, WoS is a proprietary dataset that comes with a hefty price tag for licensing. Due to licensing restrictions, we were regrettably unable to openly share our matches between firms and scientific publications with the broader research community. Fortunately, Microsoft introduced the first iteration of Academic Graph (MAG), a vast and openly available source of scientific publications, back in 2016. When the company discontinued its support for MAG in 2021, it paved the way for the emergence of the OpenAlex initiative. By embracing OpenAlex as our primary data source for scientific publications, we have overcome the previous constraints and can now include this crucial aspect of our project in DISCERN 2.0.

(b) We have used newly developed methods based on generative large language models (LLM) to match firm names to author affiliations. The Openalex dataset includes disambiguated author affiliations. However, we tested these processed affiliations and concluded that their quality is insufficient for our purposes. Instead, we have produced a direct match between the firm names and the raw affiliation strings. In the first step, we utilized batch inference using a derivative of the Llama-2 model family for entity resolution within the raw affiliation strings (see Table 1 for an example). This step greatly reduced the number of unique strings. Next, similarly to the patent matching procedure, we utilized TF-IDF methods to match firms to identified affiliations. Finally, we again used LLMs and manual reviews to validate potential matches and eliminate false positives.

Table 1: Example of Using Language Models for Entity Resolution

Raw Affiliation	**LLM Output**	**Matched Firm Name**
google inc., venice, california 90291, usa	Google Inc	GOOGLE INC
google inc., mountain view, ca, 94043, usa	Google Inc	GOOGLE INC
google inc., santa barbara, california, 93117, usa.	Google Inc	GOOGLE INC
research, google inc, mountain view, california, usa	Google Inc	GOOGLE INC
google inc (mountain view, ca, us)	Google Inc	GOOGLE INC

This table presents a sample of raw author affiliations extracted from the OpenAlex dataset. To distill organization names from these affiliations, we employed generative large language models (LLMs). Following this, we matched the extracted organizational entities with corresponding firm names through the application of TF-IDF procedures. To ensure accuracy and completeness, we then leveraged LLMs for a second round of analysis, complemented by manual verification, to validate the accuracy of the matches.

3 Preliminary Results

At this point, our data processing and matching efforts are still underway. The preliminary results include:

1. **Ultimate Owner Firms**: 8,037 firms and 11,088 associated firm names. This is a 78% increase over the DISCERN 1.0. A large part of the change is due to our decision to include R&D performing firms that do not own patents.

2. **Subsidiaries**: The subsidiary list currently includes 217,713 unique names, extracted from the SEC filings of included firms. This is a dramatic increase from DISCERN 1.0, which included approximately 50,000 subsidiary names.

3. **Patents**: The preliminary patent sample, spanning from 1980 to 2021, includes 1,923,469 patents. In comparison, the DISCERN 1.0 database, covering the period from 1980 to 2015, contained 1.34 million patents. Our current sample for the same period includes 1.41 million patents, representing a 5% increase in coverage.

4. **Scientific Publications**: The preliminary scientific publication sample includes 979,002 publications. In comparison, the DISCERN 1.0 database, covering the period from 1980 to 2015, contained 582,107 publications. Our current sample for the same period includes 762,412 publications, representing a 31% increase in coverage. The large difference in coverage is mainly due to the switch from the curated WoS dataset to the more extensive OpenAlex dataset, as well as our improved matching procedures.

4 Summary

The DISCERN 2.0 dataset represents a substantial enhancement over its predecessor, encompassing a wider time frame (1980-2021) and including a more comprehensive collection of firms, patents, and scientific publications. This version benefits from improved data quality, leveraging new sources like PatentsView and OpenAlex, and innovative matching techniques using generative large language models and enhanced TF-IDF methods. With a significant increase in data volume, DISCERN 2.0 offers a more detailed view of the U.S. corporate innovation landscape. We are committed to ongoing enhancements and maintenance of the data, making sure that it continues to be an instrumental asset for researchers studying various dimensions of corporate innovation, strategy, and economic growth.

References

Arora, Ashish, Sharon Belenzon, Matt Marx, and Dror Shvadron. 2023a. “When does patent protection spur cumulative research within firms?”, DOI: 10.1093/jleo/ewad006.

Arora, Ashish, Sharon Belenzon, and Lia Sheer. 2021a. “Knowledge Spillovers and Corporate Investment in Scientific Research,” Vol. 111, No. 3, pp. 871–898, DOI: 10.1257/aer.20171742.

———. 2021b. “Matching Patents to Compustat Firms, 1980-2015: Dynamic Reassignment, Name Changes, and Ownership Structures,” *Research Policy*, Vol. 50, No. 5, p. 104217, DOI: 10.1016/j.respol.2021.104217.

Arora, Ashish, Wesley Cohen, Honggi Lee, and Divya Sebastian. 2023b. “Invention value, inventive capability and the large firm advantage,” Vol. 52, No. 1, p. 104650, DOI: 10.1016/j.respol.2022.104650.

Arts, Sam, Bruno Cassiman, and Jianan Hou. 2023. “Position and Differentiation of Firms in Technology Space,” DOI: 10.1287/mnsc.2023.00282 Publisher: INFORMS.

Bahar, Dany, Prithwiraj Choudhury, Do Yoon Kim, and Wesley W. Koo. 2023. “Innovation on Wings: Nonstop Flights and Firm Innovation in the Global Context,” DOI: 10.1287/mnsc.2023.4682 Publisher: INFORMS.

Bessen, James. 2009. “NBER PDP Project User Documentation,” *Unpublished documentation*.

Bureau van Dijk. 2018. “Orbis Ownership Files, 2002-2015,” Bureau van Dijk, Chicago, IL. Available at `https://orbis.bvdinfo.com/`.

Center for Research in Security Prices. 2022. “CRSP Compustat, 1980-2022,” Available to Duke University through Wharton Research Data Services (WRDS).

Hall, Bronwyn H, Adam B Jaffe, and Manuel Trajtenberg. 2001. “The NBER patent citation data file: Lessons, insights and methodological tools,” NBER Working Papers: No. 8498.

Kini, Omesh, Sangho Lee, and Mo Shen. 2023. “Common Institutional Ownership and Product Market Threats,” DOI: 10.1287/mnsc.2023.4830 Publisher: INFORMS.

Papageorgiadis, Nikolaos, Andreas Procopiou, and Wolfgang Sofka. 2023. “Unintended consequences of outcome based compensation–How CEO bonuses, stocks and stock options affect their firms’ patent litigation,” *Research Policy*, Vol. 52, No. 8, p. 104816, DOI: 10.1016/j.respol.2023.104816.

Priem, Jason, Heather Piwowar, and Richard Orr. 2022. “OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts,” *arXiv preprint arXiv:2205.01833*.

Securities Data Company Platinum. 2022. “Mergers & Acquisitions Module, 1980-2022,” Refinitiv, London, GB. Provided via a Duke University subscription service.

Sheer, Lia. 2022. “Sitting on the Fence: Integrating the two worlds of scientific discovery and invention within the firm,” Vol. 51, No. 7, p. 104550, DOI: 10.1016/j.respol.2022.104550.

Standard & Poor’s. 2022. “North America Annual Compustat, 1980-2022,” Available to Duke University through Wharton Research Data Services (WRDS).

U.S. Patent and Trademark Office. 2022. “Data Download Tables,” PatentsView. Available at https://patentsview.org/download/data-download-tables.

Wharton Research Data Service. 2023. “Subsidiary Data by WRDS, 1995-2022,” Available to Duke University through Wharton Research Data Services (WRDS).