

Data Description Sheet for
“Internalizing Peer Firm Product Market Concerns:
Supply Chain Relations and M&A Activity”

Farzana Afrin
California State University Fullerton

Jinhwan Kim
Stanford Graduate School of Business

Sugata Roychowdhury
Northwestern Kellogg School of Management

Benjamin P. Yost
Boston College

Compliance with the Data and Code Sharing Policy
for the Journal of Accounting Research
Revised June 2026

The following memo provides the information requested under the Journal of Accounting Research (JAR) Data and Code Sharing Policy.

1. A description of which author(s) handled the data and conducted the analyses.

Farzana Afrin handled the cleaning of the raw datasets and the construction of the variables, assisted by two research assistants, Ziqing Wang and Taoyu Long. Farzana Afrin and Jinhwan Kim conducted the empirical analyses. All authors were involved in the research design and in interpreting the results.

2. A detailed description of how the raw data were obtained or generated, including data sources, the specific date(s) on which data were downloaded or obtained, and the instrument used to generate the data (e.g., for surveys or experiments). We recommend that more than one author can vouch for the stated source of the raw data.

The raw data were obtained between August 2019 and December 2023. The extracts used in the final analyses are dated as follows (the dates of the Wharton Research Data Services (WRDS) extracts are recorded in the dataset file names): the Compustat and CRSP data were pulled on April 10, 2022; the Compustat Segment data on November 20, 2022; the SDC mergers-and-acquisitions sample on December 17, 2022; and the He–Huang financial-institution merger list on December

19, 2022. The remaining web-based sources (the Hoberg–Phillips Data Library, PatentsView and KPSS patent data, the Zolas ALP crosswalk, and Peter K. Schott’s import data) were downloaded from their respective websites over the same period.

Farzana Afrin vouches for the stated sources of the raw data; the other authors are aware of the data sources and have assisted with decisions pertaining to data collection. The data sources and the specific data obtained from each are described below.

- **Compustat Fundamentals Annual (North America), via WRDS.** Used to construct firm-level control variables. *Variables downloaded:* gvkey, tic, cik, conm, cusip, datadate, fyear, fyr, sale, at, ib, csho, xrd, ceq, sppe, dp, che, txdb, invt, rect, cogs, dvt, oibdp, lct, prcc_f, ppent, ch, dltd, dlc, lt, act, capx, sppiv, aqc, ni, oancf, emp, and cshi (filtered to indfmt=INDL, datafmt=STD, popsrc=D, consol=C, curcd=USD), together with sic and naics industry codes from Compustat Names.
- **CRSP, via WRDS.** Used to link firms across databases, to obtain historical industry codes, and to scale institutional holdings. *Variables downloaded:* from the CRSP/Compustat Merged linking table (ccmxpf_linktable) — gvkey, lpermno, linktype, linkdt, linkenddt; from the daily stock event-names file (dsenames) — permno, siccd, naics, namedt, nameenddt; from the monthly stock file (msf) — permno, date, shrout (shares outstanding); and a CUSIP–permno link (permno, ncusip).
- **Compustat segment files and the Segment Company Names–GVKEY link table, via WRDS.** Used to build supplier–customer relationships. *Variables downloaded:* from wrds_seg_customer — gvkey, srcdate, cnms (customer name), salecs (customer sales), cid, ctype; from Seglink — gvkey, cid, srcdate, cgvkey (customer gvkey), ctic, ccusip, cconm, cnms, scusip.
- **Mergers and acquisitions, via Refinitiv (Thomson) Securities Data Company (SDC) Platinum.** *Deal-level variables downloaded:* master_deal_no (deal identifier), dateann, dateeff, datewith (announcement, effective, and withdrawal dates), acusip and tup (acquirer and target CUSIPs), amanames and tmanames (acquirer and target names), anation and tnation, apublic and anationcode, form (deal form), status, pheld, pctown, pctacq (percent held, owned, acquired), mv (deal value), pct_cash, pct_stk, attitude, and ATF_MID_DESC and TTF_MID_DESC (acquirer and target mid-industry descriptions).
- **Institutional ownership, Thomson Reuters Institutional (13F) Holdings, via WRDS.** Combined with CRSP shares outstanding to compute institutional ownership. *Variables downloaded:* cusip, mgrno, mgrname, rdate, shares, stkname, ticker.
- **Product similarity and industry concentration, via Hoberg–Phillips [website](#).** *Variables obtained:* gvkey, year, tnic3tsimm (text-based total similarity score), and tnic3hhi (TNIC-based Herfindahl index / industry concentration).

- **Product market fluidity**, via Hoberg–Phillips [website](#). *Variables obtained*: gvkey, year, prodmktfluid.
- **Vertical relatedness**, via Gerard Hoberg’s [website](#). *Variables obtained*: gvkey1, gvkey2, year, vertscore.
- **Patents**, via PatentsView and the github [website](#) hosting the data for Kogan, Papanikolaou, Seru, and Stoffman (2017, *Quarterly Journal of Economics*). *Variables obtained*: from the KPSS data (KPSS_2019_public.csv and KPSS_2020_public.csv) — patent_num, permno, filing_date, issue_date, cites (forward citations), xi_real (stock-market-based patent value); from PatentsView (patent.dta and PatentsView_KPSS_combo_yr_figs.dta) — patent id, grant date, and cpc_group_id (CPC technology class).
- **Technology-class-to-industry crosswalk**, via Nikolas Zolas' Patent [Crosswalk](#). Used to map patents from technology classes to NAICS industries. *Variables obtained*: cpc4 (4-digit CPC class), naics (4-digit NAICS, 1997/2002/2007 vintages), and probability_weight.
- **Financial-institution mergers, constructed from the Appendix of He and Huang (2017, Review of Financial Studies)**. Used to construct the common-ownership shock. *Variables obtained*: Target_Name, Acquiror_Name, Target_Mgrno, Acquirer_Mgrno, Announcement_Date, and Completion_Date.
- **Joint ventures and alliances**, via Refinitiv (Thomson) SDC Platinum, Joint Ventures and Alliances. *Variables obtained*: participant CUSIPs (CUSIP1–CUSIP21), begin_year, and end_year.
- **Redacted disclosures and trade secrets, hand-collected from firms’ SEC filings on EDGAR**. (<https://www.sec.gov/edgar>) *Variables (customer gvkey–year level)*: cgkey, year, a trade-secret indicator (ctrade_secret), and a redaction indicator (credaction).
- **Tariffs, product-level U.S. import data from Peter K. Schott’s International Trade Data**. Used to compute ad valorem tariff rates and identify tariff cuts. *Variables obtained (by SIC industry–year)*: sic, year, customs (customs value of imports), cif, and duties (duties collected).

3. If the data are obtained from an organization on a proprietary basis, the authors should privately provide the editors with contact information for a representative of the organization who can confirm data were obtained by the authors. The editors would not make this information publicly available. The authors should also provide information to the editors about the data sharing agreement with the organization (e.g., non-disclosure agreements, and any restrictions imposed by the organization on the authors). In particular, the authors should indicate if an organization or data provider imposes restrictions on the publication of the results, has not given the authors full control of the relevant data, requires

that the results must be reviewed or approved prior to public release of the paper or publication.

All data used in this study are obtained either from commercial databases accessed through institutional licenses with the data providers named above or hand-collected from publicly accessible documents (e.g., SEC EDGAR). No data were obtained on a proprietary basis. No data provider imposes any restrictions on the use of the data or on the publication of the results, and the authors retain full control of all relevant data.

4. A complete description of the steps necessary to download, obtain or collect as well as process the data used in the final analyses reported in the paper. For experimental and survey papers, we require information about the instructions and instruments used to generate the data, subject eligibility and/or selection, as well as any exclusion criteria. The full set of instructions and instruments can be provided in the online appendix.

The steps to collect and process the data used in the analyses are described in Sections 3, 4, and 5 and the Online Appendix of the manuscript. The readme file accompanying our code provides additional details. The main steps are:

- We combine the Compustat Customer Segment data with Compustat Fundamentals Annual, using the Compustat Customer Segment Company Names–GVKEY link table, to construct a dataset at the supplier–customer–year level, retaining supplier–customer–years with non-missing identifiers and sales data.
- We identify suppliers' major customers (those accounting for 10% or more of a supplier's total revenues) using the Compustat Segment files, matching each customer's name to the correct gvkey via the Seglink file and retaining only U.S. public-firm customers (excluding private, foreign, or government customers and those whose identity is not revealed). We then obtain customers' total product similarity scores from Hoberg and Phillips (2016) to proxy for customers' competition concerns. Because the analyses are at the supplier–year level, we collapse the data to one observation per supplier–year and construct a sales-weighted sum of the product similarity scores of all of a supplier's customers. This yields 34,128 supplier-years from 1989 to 2014, which we match to suppliers' M&A transactions from 1990 to 2019.
- Data on suppliers' acquisitions come from Refinitiv SDC Platinum. We retain deals announced from 1990 to 2019 classified as a merger, acquisition of majority interest, or acquisition of assets, and require that the supplier owns less than 50% of the target before and more than 90% after the deal (Bena and Li 2014). In the main tests we focus on vertically conflicted transactions — those in which the target is a customer's rival or a supplier to a customer's rival, where rivals share the customer's 3-digit SIC code. Control variables are constructed from Compustat. The final sample consists of 22,869 supplier firm-year observations for 3,683 unique firms, during which these suppliers conduct 284 vertically conflicted transactions (the primary dependent variable).

- In additional tests we combine the sample with data from Refinitiv SDC (Joint Ventures and Alliances), Thomson Reuters 13F filings, firms' SEC filings on EDGAR, the Hoberg–Phillips Data Library (similarity, fluidity, and vertical relatedness), the Zolas et al. CPC-to-NAICS patent crosswalk, PatentsView and KPSS (2017) patent data, the financial-institution mergers from He and Huang (2017), and product-level U.S. import data.

5. After downloading or obtaining the raw data, all manipulations of the data should be done via computer programs. The code for these manipulations should be included in the code submitted upon acceptance (see below). No manipulations of raw data can take place manually or outside the computer code provided. If compliance with this requirement is not feasible, the authors need to explain and disclose any manipulations of the raw data (e.g., manually created variables or file conversions). When feasible, we also encourage the authors to share the code that downloads the data.

All manipulations of the raw data are performed via computer programs written in SAS and Stata; no manipulations were performed manually or outside the provided code. The individual programs that download the data, construct the variables, and produce the results are listed and described under Question 6 below. The only data collected by hand are the customer trade-secret and redaction indicators gathered from firms' SEC filings on EDGAR; all subsequent processing of these data is done in code.

6. The computer programs (i.e., code) used to (1) convert the raw data into the final dataset used in the analysis, (2) to execute the statistical or econometric analysis, and (3) to generate the tables or to produce the output used in constructing tables of the manuscript. A brief description that enables other researchers to understand and run the code should be provided. The purpose of this requirement is to facilitate replication and to help other researchers understand in detail how the raw data were processed, the final sample was formed, variables were defined, outliers were treated, and which commands were used in the analysis, etc. This code or programming is in most circumstances not proprietary. However, we recognize that some parts of the code or data generation process may be proprietary, including from the authors' perspective. Therefore, instead of disclosing the proprietary portion of the code or program, researchers can provide a detailed step-by-step description of the code or the relevant parts of the code such that it enables other researchers to arrive at the same results that the authors obtained and presented in their manuscript. In such cases, the authors should inform the editors upon initial submission, so that the editors can consider an exemption allowing the step-by-step description. Whenever feasible, authors are required to provide the identifiers (e.g., CIK, CUSIP) for their final sample. Authors should consult our FAQ Sheet on the JAR website for further details.

The replication package includes annotated SAS and Stata code, to be run in the following order, together with a readme file (Readme.txt) that describes each file:

- **“0_Get_raw_data.sas”** — downloads and assembles the raw Compustat, CRSP, Compustat Segment, SDC mergers-and-acquisitions, and Thomson Reuters 13F data from WRDS.

- **“1_X_var_construction.do”** — constructs the measures of customers’ competition concerns and the supplier firm-year independent variable of interest.
- **“2_Y_var_construction.sas”** — identifies customer rivals (by 3-digit SIC code) and the suppliers of those rivals, and flags the M&A transactions that are vertically conflicted (the dependent variable).
- **“3_Controls&CS_vars_construction-Part 1.sas”** — builds the control and cross-sectional variables (joint ventures, relationship length, R&D, patents, trade secrets, and redaction), the customer-switch and common-ownership datasets, and merges them into the supplier firm-year dataset.
- **“3_Controls&CS_vars_construction-Part 2.do”** — builds the technology-class-to-NAICS crosswalk and technological-similarity measures, the customer-level controls, the vertical-relatedness dataset, the rival-weighted customer competition-concern measure, and the tariff-cut dataset.
- **“4_Final_dataset_construction.do”** — merges all variables into the main dataset and renames, log-transforms, winsorizes (at the 1st and 99th percentiles), and drops missing values to produce the final estimation sample.
- **“5_AKR_regressions.do”** — executes all statistical analyses and generates the tables and figures reported in the manuscript.

We also provide two files containing the identifiers for our main sample:

- **“dealids.dta”** — the SDC master deal numbers (deal_id) for the 284 vertically conflicted transactions in our main sample; and
- **“firmids.dta”** — the Compustat gvkeys (sgvkey) and company names (conm) for the 3,683 unique supplier firms in our main sample.

7. A comprehensive log file that shows the execution of the entire code. This log file should cover all the steps that convert the raw data into a final dataset and the execution of all statistical and econometric analyses presented in the tables of the manuscript. The portion of the log file that shows proprietary code or data may be masked. In this case, the reader should be referred to the step-by-step description provided as per the requirements in Item 6.

The replication package includes log files documenting the execution of the code: the construction of the variables (log_1.txt, log_2.smcl, and log_3.txt), the assembly of the final dataset, and all statistical analyses, tables, and figures reported in the manuscript (log_4.smcl, which includes the execution of 4_Final_dataset_construction.do and 5_AKR_regressions.do).

8. An assurance that the data and programs will be maintained by at least one author (usually the corresponding author) for at least six years, consistent with National Science Foundation guidelines.

The authors agree to collectively maintain the data and programs for at least six years, consistent with National Science Foundation guidelines.