

Data Description Sheet
“Common Investor Relations Representation”

David Volant

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

David Volant handled the data and conducted the analysis.

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 is able to vouch for the stated source of the raw data.

The data were retrieved between 2021 and 2025. I describe below the data sources and the particular data retrieved from each:

Earnings-announcement press releases: Retrieved from the WRDS SEC Analytics Suite on November 11, 2021. The text of 8-K earnings-announcement exhibits, with HTML markup removed, is used to identify external investor-relations (IR) firms from the investor-contact email domains that appear in the releases.

Compustat: Quarterly and annual Compustat data retrieved from WRDS on August 16–20, 2022. Downloaded variables include gvkey, data date, fiscal year, income before extraordinary items, total assets, sales, common shares outstanding, price, stockholders’ equity, long-term debt, SIC code, stock exchange, and auditor.

CRSP: Daily and monthly stock data retrieved from WRDS in August 2022. Downloaded variables include permno, date, holding-period return, price, shares outstanding, and the CRSP name and CUSIP history.

I/B/E/S: Analyst EPS forecast detail and management guidance detail retrieved from WRDS in August 2022. Downloaded variables include ticker, analyst code, forecast and announcement dates, forecast period-end date, and the guidance metric and value.

Institutional ownership: Thomson-Reuters institutional (13-F) holdings retrieved from WRDS in August 2022. Downloaded variables include manager number, CUSIP, report and file dates, and the number of reporting institutions.

SEC filings: 10-K and 10-Q filing dates and 8-K Item 5.02 filings retrieved from the WRDS SEC Analytics Suite (2021–2022).

Product-market similarity: Text-based Network Industry Classification (TNIC) product-similarity scores from the Hoberg-Phillips Data Library (<https://hobergphillips.tuck.dartmouth.edu/>).

News and press-release counts: RavenPack Dow Jones Edition, accessed through WRDS.

Reference data: Fama-French 48-industry classifications from Kenneth French's Data Library; Russell 2000 index membership from CRSP; and merger firm-years from SDC Platinum.

Linking files: The I/B/E/S–CRSP and CRSP–Compustat linking tables from WRDS are used to map identifiers across databases.

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, any restrictions imposed by the organization on the authors, such as restrictions to publish certain results). 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.

The commercial databases described above are accessed and licensed through the institutions where I conducted this research, including the University of Iowa (where the dissertation was completed) and Indiana University. No data provider imposed restrictions on the publication of the results, required that results be reviewed or approved prior to public release, or withheld the author's full control of the relevant data.

Publicly available reference data are cited in Item 2: the Fama-French industry definitions (Kenneth French's Data Library) and the TNIC product-similarity data (Hoberg-Phillips Data Library, <https://hobergphillips.tuck.dartmouth.edu/>).

4. A complete description of the steps necessary to collect and 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 complete sequence of steps used to collect and process the data is described in Sections 3 and 4 of the manuscript and in Appendix A (variable definitions). The replication package reproduces every step: a Python pipeline builds the analysis datasets from the source databases (see Item 6), and the package README summarizes the data sources, the order of the construction steps, and the resulting datasets. In brief, external IR firms are identified from earnings-announcement press releases; firms that outsource IR are assembled into a firm-quarter sample; treated firm pairs (two firms that begin sharing an external IR firm) are each matched to control pairs in the same industry and size decile and stacked in a three-year pre/post event window; and commonality measures (common ownership, common analyst coverage, common

guidance, return comovement, and same industry / size / valuation / profitability / auditor indicators) are computed for each firm-pair-quarter.

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.

After the raw data were obtained, all manipulations of the data are performed by the computer programs included in the replication package; no manipulations were performed manually or outside the provided code. The package includes a Python data-construction pipeline, the Stata do-files that produce the analyses and tables, and a README describing the execution environment and directory structure. The one non-database input, the identification of external IR representation from press releases, is itself produced by code (see Item 6).

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 replication package contains the following:

(1) Data construction (Python). The python/ folder converts the source database extracts into the five analysis datasets used in the paper: the main firm-pair-quarter panel and the textual-analysis, economic-characteristics, dissipation, and determinants datasets. The script run_all.py executes every construction step in order (modules s00 through s11), each documented with the operation it performs, and writes a timestamped execution log. A README describes how to obtain the database snapshots and run the pipeline.

(2)–(3) Analysis and output (Stata). The stata/ folder contains the do-files that read the five datasets and produce the published tables.

Text-derived measures. Two measures are constructed from earnings-announcement press-release text: the identification of external IR representation (the treatment) and the disclosure-text cosine similarity. The code for both is provided in the source_measures folder of the replication package (external_ir_identification and disclosure_text_similarity); it documents how the two measures are constructed from the press-release text. The data-construction pipeline runs from the resulting firm-level IR assignment and the finished text-similarity file, and is provided in full.

Identifiers. I provide an identifier file for the final sample, listing firm-level keys (GVKEY / PERMNO) for the sample firms. The external IR identifiers are redacted and not provided within this data package.

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 package includes execution logs covering the path from the source database extracts to the final datasets and to the published tables. The Python pipeline writes a timestamped log of every construction step (`code_execution_log.log`), beginning with the parse of the earnings-announcement press releases that identifies external IR representation and continuing through to the analysis datasets; the Stata analysis produces a log of the regressions and table generation.

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.

David Volant will maintain the data and programs for at least six years, consistent with National Science Foundation guidelines.