

Data Description
“Quid Pro Quo? Private Information Flows in Shareholder Activism: Evidence from Mutual Fund Families”

By Eunjee Kim and Hai Pham

This data description sheet is provided to fulfill the requirements of the *Journal of Accounting Research* data policy for submission of a manuscript.

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

All co-authors of this paper were involved with handling various aspects of the data collection, verification, and analysis. The assembly and merging of data files to create the main dataset, as well as the empirical analyses, were conducted primarily by Hai Pham.

2. *A detailed description of how the raw data were obtained or generated, including data sources, the 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.*

We obtained our data from six major public databases, including (1) activist campaigns from the FactSet corporate activism (SharkRepellent) database; (2) firm-level financial data from the Compustat database; (3) stock return data from the Center for Research in Security Prices (CRSP); (4) mutual fund data from CRSP Mutual Fund and Thomson Reuters Mutual Fund holding (S12) databases; (5) institutional holdings data from the Thomson Reuters S34 file; and (6) voting results and mutual fund vote data from ISS Voting Analytics (Form N-PX) databases. All data were downloaded between August 1, 2021 and March 31, 2025. The details of the datasets are as follows:

- Shareholder activism: We obtain the list of activist campaigns spanning from 1995 to 2020 from the FactSet corporate activism (SharkRepellent) database. The dataset was obtained in August 2021.
- Firm-level financial data: We obtained firm-level financial data from Compustat. The firm-quarter level data was obtained from Compustat Fundamental Quarterly on May 5, 2023, and the firm-year level data was obtained from Compustat Fundamental Annual on March 2, 2022. All data were retrieved via Wharton Research Data Services (WRDS) using SAS.
- Stock returns: These data were obtained from CRSP on April 26, 2023, via WRDS using SAS.
- Mutual fund data: We obtained detailed mutual fund information, such as name, fund flows, and total net assets, from the Center for Research in Security Prices (CRSP) Survivor-Bias-Free US Mutual Fund database. We then obtained portfolio holdings data from the Thomson Reuters Mutual Fund holding database (S12) and merged the ownership data with the CRSP Monthly Stock File based on the MFLINKS files. Because our main analysis focuses on actively managed US equity mutual funds, we exclude index funds using the index flag code and fund names provided by CRSP. To construct the data at the mutual fund family level, we hand-matched mutual fund family names to address missing identifiers. Data are retrieved via WRDS using SAS, with downloads conducted between August 1, 2021, and September 10, 2021.

- Institutional holdings data: We obtain institutional holdings data from the Thomson Reuters S34 database, downloaded in August 2024.
 - Voting results and mutual fund vote: We obtained voting results and mutual fund votes from ISS Voting Analytics (Form N-PX) databases. Because ISS Voting Analytics and S12 lack common institution identifiers, we use fuzzy matching and hand-matched fund family names for our analyses. Voting data were retrieved between September 2019 and March 2025.
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.*

All data used in this paper are retrieved from public databases, as detailed above.

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.*

We describe our data in Section 3.1 and provide further details in item (5) below.

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 data were done via computer programs.

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.

All data manipulation and dataset construction were performed primarily using SAS and Stata. The final analyses were conducted in Stata. Further details on sample construction and variable definitions are provided in Section 3.1. Below, we also provide code in accordance with JAR's Data and Code Sharing Policy, including the program files used to assemble the analysis datasets and generate the main results.

- `step1a_makedepvar.sas`, `step1b_makedepvar_index.sas`, `step1c_makedepvar_posnegcar.sas`, `step1e_makedepvar_topx.sas`, `step1f_makedepvar_fund.sas`, `step1g_makevote.sas`, `step1h_makevotenpx_part1.sas`, `step1h_makevote_part2.sas`, `step1i_makedepvar_aum.sas`: The SAS programs used to construct dependent variables for different samples used in our paper.
- `step2a_assembledata_main.sas`, `step2b_assembledata_index.sas`, `step2c_assembledata_negcar.sas`, `step2d_assembledata_poscar.sas`, `step2e_assembledata_topx.sas`, `step2f_assembledata_fund.sas`, `step2g_assembledata_vote1.sas`, `step2g_assembledata_vote2.sas`, `step2a_assembledata_aum.sas`: The SAS programs used to assemble datasets and construct samples used in our paper.
- `Step3_makesamples.do`: the Stata program used to create final samples.
- `Step4_regressions.do`: the Stata program used to run the analyses.
- The file “ids.dta” provides the company identifiers (GVKEY) of the firms in the final dataset (3,387)

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.*

We provide log files documenting the full execution of the code.

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.*

Hai Pham will maintain the programs and dataset for at least six years following publication.