

**Datasheet for**

**“How Does Judges’ Personal Exposure to Financial Fraud Affect White-Collar Sentencing?”**

Trung Nguyen, Aneesh Raghunandan, and Alexandra Scherf

## **1. Authors**

All three authors handled the data collection process, including gathering the data and merging it together to create the final estimation datasets, and can vouch for all sources of data. Many data-cleaning and merging tasks, related to the court case data as well as data on judges' holdings, were done by research assistants under the authors' direct supervision and the authors regularly spot-checked and audited their output.

## **2. Data Sources**

We outline below the source for each dataset we used. Judge holdings data were obtained in Summer 2021 and Spring 2022. Fraud data used to identify treatment observations for our analyses were obtained in Spring and Summer 2021. Court case data used to construct our dependent variables were obtained in Summer 2021 and Summer 2022. Market returns data from CRSP were obtained in Summer 2021, Summer 2022, and again in Spring 2023. Compustat data were obtained in Summer 2021.

### *2.1 JudicialWatch*

We obtained initial data on judges' personal financial disclosures from JudicialWatch. We obtained files as PDF documents, which were generally not machine-readable; these files therefore required extensive use of optical character recognition (OCR), for which we used ABBYY software. We then used fuzzy matching to merge this data with firm identifiers (e.g., GVKEY) and hand-checked all matches for accuracy.

### *2.2 Free Law Project*

To supplement the JudicialWatch data, we obtained additional information on judges' personal financial disclosures from the Free Law Project, hosted by CourtListener ([www.courtlistener.com](http://www.courtlistener.com)). We then used fuzzy matching to merge this data with firm identifiers (e.g., GVKEY) and hand-checked all matches for accuracy.

### *2.3 AAER Data*

We obtained the AAER database from the authors of Dechow et al. (2012), in spreadsheet form, which we then merged with the rest of our data. We hired research assistants to manually identify, using Factiva as well as Google searches, the first date for which each case first publicly came to light; other than these dates, no data was manually entered.

### *2.4 Stanford Securities Class Action Lawsuit Database*

We obtained the SSCAC data in spreadsheet form from Stanford Law School, which contained identifiers enabling a merge to the rest of our data.

### *2.5 Violation Tracker*

We obtained additional financial misconduct information from Violation Tracker, which we merged into the rest of our data. We hired research assistants to manually identify, using Factiva as well as Google searches, the first date for which each financial misconduct case derived from Violation Tracker within our sample first publicly came to light; other than these dates, no data was manually entered.

#### *2.6 Federal Judicial Center*

We collected two datasets from the FJC: one on civil case outcomes, and one on judges' biographical information. We downloaded these datasets directly from WRDS as .csv and .dta files.

#### *2.7 Corporate Prosecution Registry (CPR)*

We obtained data on criminal prosecutions from the Corporate Prosecution Registry (CPR). We obtained this data in spreadsheet form.

#### *2.8 PACER*

To supplement the Corporate Prosecution Registry data, we obtained data from PACER that included information on the judges hearing cases.

#### *2.9 CRSP*

We obtained CRSP data to construct short- and long-window returns variables. We downloaded Stata .dta files directly from WRDS.

#### *2.10 Compustat*

We obtained Compustat data for the purposes of fuzzy matching our other datasets as well as to construct defendant-specific control variables. We downloaded Stata .dta files directly from WRDS.

### **3. Proprietary Data**

All of the data are publicly available from the sources listed above. Many of these sources require a paid subscription, but our understanding is that any researcher who wishes to purchase any of the data may do so.

### **4. Data Processing**

These steps are described throughout the paper, primarily in Section 4, and provided in the code outlined below.

### **5. Code**

We enclose several Stata and R code files pertaining to the data as well as corresponding Stata log files.

- *2. Fraud\_universe\_FU\_data.do* and *2.1. Fraud\_universe\_FU\_data\_redating.do* creates the sample of fraud events for the paper with manual re-dating.
- *3. Financial\_disclosure\_FD\_data.do* cleans judges' holdings obtained from Personal Financial Disclosures (PFD) to link investments to GVKEYs. This file relies on earlier R code written in *FinancialDisclosure\_clean\_company\_names\_8.23.21\_final.R*.
- *4. Courtlistener\_CL\_data.do* processes judge financial disclosure data from CourtListener prior to taking the union with the judges' investments obtained from financial disclosure data (from *3. Financial\_disclosure\_FD\_data.do*).
- In two do-files, *5. Comprehensive\_data\_merge\_FD\_CL.do* as well as *5.1. Comprehensive\_data\_merge\_FD\_CL\_detail.do*, we take the union to obtain a comprehensive judge-year holdings data. We prepare the unified dataset for later merging with case data using judge names in *6. Comprehensive\_data\_clean.do*.
- We next use the full dataset of judges' investments to identify judges' indirect investments in mutual funds in *7. Match\_judges\_indirect\_holdings\_to\_CRSP.do*.
- We identify judges' exposure to fraud through these indirect investments in *7.1. Identifying\_indirect\_fraud\_exposure\_and\_buysell\_adjust.do*, adjusting for buy and sell activity to ensure the judge held the investment at the time of the fraud. Analogously, *8. Identifying\_direct\_fraud\_exposure\_and\_buysell\_adjust.do* identifies judges' direct exposures to fraud. We include short-term market reactions (CAR) for each exposure event which are calculated in *11. Short-term\_market\_reactions\_for\_fraud\_events.do*.
- *9. Define\_treatment\_and\_control\_groups.do* defines treatment and control judges based on their exposures to fraud.
- *10. Case\_data\_CD.do* loads the civil and criminal case data, calculates outcome variables, and prepares it for merging with the unified data of judges' investments.
- *12. Control\_variables.do* and *12.1 Control\_variables\_continued.do* create the control variables as outlined in the paper.
- *13. Prepare\_data\_for\_tables.do* combines the treatment data, case data, and control variables to create the dataset used for paper's analyses.
- *14. Create\_tables.do* executes the analyses for the paper.
- *15. Untabulated\_analyses\_and\_JAR\_memo.do* executes the untabulated and supplemental analyses for the paper.

We also attach a list of all distinct judge identifiers (names) used in the paper's primary sample.

## **6. Data and Code Retention**

We commit to retaining all data and code for at least six years, in line with NSF guidelines.