

# Does U.S. immigration policy facilitate financial misconduct?

**Ruiting Dai**

*Drexel University*

[dan.dai@drexel.edu](mailto:dan.dai@drexel.edu)

**Xuanjun Dong**

*Shanghai University of Finance and Economics*

[dongxuanjun@mail.shufe.edu.cn](mailto:dongxuanjun@mail.shufe.edu.cn)

**Nemit Shroff**

*Massachusetts Institute of Technology*

[shroff@mit.edu](mailto:shroff@mit.edu)

**Qin Tan**

*City University of Hong Kong*

[qintan@cityu.edu.hk](mailto:qintan@cityu.edu.hk)

## Compliance with Data Policy for the Journal of Accounting Research

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

Ruiting Dai, Xuanjun Dong, and Qin Tan are jointly responsible of data handling and analyses.

### *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.*

We downloaded the Labor Condition Application (LCA) submission records for 2001 to 2019 from the Department of Labor (DOL) on July 4, 2022. At that time, the data from 2001 to 2007 was provided by The Foreign Labor Certification Data Center owned by the Department of Labor (DOL) via <https://www.flcdatcenter.com/caseh1b.aspx> and the data from 2008 onwards was provided via <https://www.dol.gov/agencies/eta/foreign-labor/performance>, so we used both links to download the data.

We downloaded the H-1B visa lottery approval data on Aug 14, 2022 from <https://www.uscis.gov/tools/reports-and-studies/h-1b-employer-data-hub/h-1b-employer-data-hub-files>.

We downloaded firm-level financial data from Compustat on May 26, 2022.

We downloaded restatement data from Audit Analytics on Dec 5, 2022.

We obtained investor portfolio data from Thomson's 13-F database on Jul 8, 2020.

We obtained the analyst following data from IBES on Feb 25, 2022.

We downloaded foreign subsidiaries data from Professor Dyreng's website ([https:// sites.google.com/site/scottdyreng/Home/data-and-code](https://sites.google.com/site/scottdyreng/Home/data-and-code)) on Mar 8, 2023.

We obtained firms' historical headquarters location from 10-K header data provided by Professor McDonald (<https://sraf.nd.edu/data/>) on Mar 8, 2023.

We obtained the degree of non-compete provisions from Aobdia (RAST 2018) on Mar 2, 2023.

We obtained top-15 most dependent cities based on Table 1 of Kerr and Lincoln (JLE 2010).

We obtained per capita income at state level from <https://www.bea.gov/data/income-saving/personal-income-by-state> on Oct 6, 2022.

We obtained permanent resident data from <https://www.dhs.gov/immigration-statistics/readingroom/LPR/LPRcounty> on Sep 28, 2022.

We downloaded the internal control weakness data from Audit Analytics on Mar 11, 2024.

We downloaded the employee concerns / strengths from MSCI on Mar 11, 2024.

We obtain data on employee whistleblower complaints filed with OSHA from Call, Martin, Sharp, and Wilde (2018) on July 6, 2023 and then supplement these data with mentions of employee whistleblowers in media articles obtained from Call, Kedia, and Rajgopal (2016) on July 3, 2023.

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 have to be reviewed or approved prior to public release of the paper or publication. This information should be provided to the editors upon submission.*

Our paper does not use proprietary data.

*To be provided in the paper or the online appendix:*

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 provide a complete description of the steps used to create the sample including sources of raw data in Sections 5 and 6. All variables are as described in the body of the paper, Appendix A and page 12 of the Online Appendix.

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 are performed via computer programs except the following four tasks (also marked in the code files we submitted):

1. After using Python to perform the name matching between employer names in LCA files and firm names in Compustat, our research assistants (RAs) manually inspect each match to ensure accuracy.
2. Our RAs manually go through each job title in LCA files to identify those that are related to accounting, engineering, and human resources, as well as the job titles for managerial positions.
3. Since the US city names are spelled or abbreviated differently across datasets, we manually match cities heavily dependent on immigrants as identified in Kerr and Lincoln (2010) with firms' headquarter cities in Notre Dame dataset.
4. We manually fill in the missing coordinates of the working cities in LCA files and firms' headquarter cities to calculate the variable named *DISTANCE TO HEADQUARTER*.
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.*

We submitted a step-by-step code used to perform the three tasks listed in item 6. We also described the purpose of each step in ReadMe.doc (included the code package). Firm identifiers (GVKEY) for our final sample are listed in firm identifiers.xls.

- 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 submitted a comprehensive log file package that tracks the execution of each step of the code. For example, the log file named “P2\_Independent Variable of Interest” corresponds to the Stata code named “P2\_Independent Variable of Interest”. When we occasionally use SAS to merge datasets or calculate certain variables, we clearly marked which portion of code is SAS in Stata code and provided the log file from running the SAS code as a separate log file (e.g., “P3\_Main Dependent Variables sas log”).

- 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.***

Ruiting Dai, Xuanjun Dong, and Qin Tan will maintain the data and programs for at least six years.