Data Description Sheet for "Financial Gatekeepers and Investor Protection: Evidence from Criminal Background Checks" by Kelvin Law and Lillian Mills (March 11, 2019)

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

Kelvin Law handled the data and conducted the analyses. Kelvin Law and Lillian Mills contribute equally to the writing of the paper.

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.**

Kelvin Law assigned four student assistants to use Python to scrape all data publicly available from BrokerCheck of the Financial Industry Regulatory Authority (FINRA) during five months from May to September 2017. 1,307,858 raw json files for financial advisors were obtained in August 2017, and 16,242 raw json files for advisory firms were obtained in September 2017. This sample is comparable to 1.2 million records used in the study of Egan, Matvos, and Seru (2017a). Kelvin Law also independently re-scraped 2,898 files to ensure the scraping is complete and accurate. All files were then parsed and merged into csv files in Python. Three student assistants were assigned in parsing the raw files, and their files were cross-checked to ensure files are parsed correctly. Kelvin Law also manually went through 50 files to ensure that no data are omitted during parsing. Census-related data are from the American Community Survey in Social Explorer. Mutual fund data are from CRSP Mutual Fund database in WRDS. Google search data are from Google Trend. List of Madoff's victims is collected from court documents released by the U.S. federal bankruptcy court.

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).**

Not applicable.

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.**

All relevant steps are detailed in *Section 3 Sample Data* in the manuscript and outlined in the computer programs described below.

5. **The computer programs or code used to convert the raw data into the final dataset used in the analysis plus a brief description that enables other researchers to use this program. 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, 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 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 final dataset used in the analysis. In such cases, the authors should inform the editors upon initial submission, so that the editors can consider an exemption from the code sharing requirement. Whenever feasible, authors should also provide the identifiers (e.g., CIK, CUSIP) for their final sample. Authors should consult our FAQ Sheet on the JAR website for further details.**

   "*LawMills 2019 JAR Replication Codes Part 1 Convert Raw Data.do*" file details the steps from converting the raw FINRA data into the main samples used in this paper. "*LawMills 2019 JAR Replication Codes Part 2 Estimate Regressions.do*" file replicates Tables 1-10 in the paper. The file "*Identifiers for Final Sample*" contains the identifiers (i.e., CRD) of financial advisors in our sample.

6. **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.**

Both authors will maintain a shared Dropbox account with the complete set of code and data that replicate all main tables for at least six years.