Data Description Sheet for the paper "**The disciplinary effect of social media: Evidence from firms' responses to Glassdoor reviews**"

(June 2021)

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

Both authors handled the data and conducted the 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.

All data used in this paper is archival data, obtained from commercially available sources.

The first major dataset used in the paper is CSR scores obtained from MSCI ESG KLD STATS database (henceforth, KLD). The KLD data used in the current version of the paper was downloaded from WRDS through institution subscription in December 2020, covering firms from 1991 and 2018. We start the main sample period from 2003 when KLD increases coverage to include companies in MSCI KLD 400 Social Index and 1000 Largest US Companies. We match KLD to CRSP based on historical CUSIP and manually check observations with irregular CUSIP in KLD. We further match the data to Compustat, IBES, and Thomson Reuters to obtain common firm characteristics (e.g., size, leverage, financial performance, analyst following, and institutional holdings). We then exclude firms with missing values on these characteristics. To center the analysis around the launch of Glassdoor.com, we require firms to have at least observation before and after 2008.

The second major dataset used in the paper is Glassdoor reviews obtained from Glassdoor. We collected Glassdoor review data from March 2019 to November 2019. We contacted Glassdoor in March 2019 and obtained the written permission to use all publicly available data on Glassdoor from Glassdoor's chief economist Andrew Chamberlain with no strings attached to the publication of our results. We first identify each public firm in Glassdoor by manually inputting its name into the search engine on Glassdoor.com and pick the Glassdoor URL based on company name, logo, and industry if multiple matches are shown in the search results. During the manual match, we also collected the date and rating of the first review. In August 2019, we hired X-Byte Enterprise Crawling to collect the numerical ratings and narratives of all Glassdoor reviews for sample public firms. We got the final batch of data in November 2019.

To measure CSR disclosures, we used the data from 10-K filings (mentioned in the footnote) and CSR reports (main CSR disclosure measure used in the paper), respectively. The 10-K filings were downloaded from Loughran-McDonald's website (Stage One 10-X Parse files) in February 2019.[1]

[1] https://sraf.nd.edu/data/

CSR reports were purchased from Corporate Register and the data includes all CSR Reports from 2000 through 2018 collected as of January 13th, 2021.

We used BoardEx to identify CEO turnover and CEO age. The BoardEx data was downloaded via NYU's institution subscription in May 2019.

Both authors have access to the data described unless noted explicitly.

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., nondisclosure 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.*

A complete description of the steps used to construct the sample is described in Section 3.1 of the paper. The raw data source is also described in Section 3.1. Calculations of all variables are described in Section 3.2 of the paper and Appendix B.

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.*

Please see the attached replication code. Included in this are:

- A README file with descriptions about how to use the replication code.

- Two sets of Stata codes that reproduce the analyses (tables & figures) reported in the main body of the paper and the Appendix.

- A Stata dataset (identifiers.dta) containing our sample firms (gvkey and year).

In addition to the above, we provide detailed descriptions in the paper, and our Appendix, on the steps taken for our data construction and variable creation.

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.*

The authors agree to maintain the data and programs used in this paper for six years.