

“Megadeal” subsidies, local spillovers, and corporate innovation

Data Description Sheet

June 30th, 2026

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

Co-author Yoojin Lee handled collecting "Megadeals" subsidies data from Good Jobs First (GJF). Co-authors Shaphan Ng and Aruhn Venkat jointly handled corporate patent data from various sources. All co-authors are involved in conducting the analyses throughout the paper to varying degrees.

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 use data from the following sources:

1) Corporate patent data

- a. We downloaded corporate patent data from Kogan et al. (2017)'s repository (<https://github.com/KPSS2017/Technological-Innovation-Resource-Allocation-and-Growth-Extended-Data>) on May 30, 2024.

2) Inventors' location data

- a. We downloaded the inventor location information from USPTO PatentsView.org (<https://patentsview.org/download/data-download-tables>) on October 19th, 2023.

3) Corporate Megadeal subsidies

- a. We downloaded the Megadeal subsidies from Good Jobs First (GJF) (<https://subsidytracker.goodjobsfirst.org/>) on November 27, 2019. We supplement this data with hand-collected county information conducted from May to October in 2021.

4) County's macroeconomic data

- a. We obtained county unemployment rates from the Bureau of Labor Statistics (<http://www.bls.gov/lau/data.htm>). We obtained this data on October 11, 2021.
- b. We obtained county income per capita from the Bureau of Economic Analysis (<https://www.bea.gov/itable/regional-gdp-and-personal-income>). We obtained this data on November 1, 2023.
- c. We obtained county industry employment location quotient from the Bureau of Labor Statistics (<https://www.bls.gov/cew/downloadable-data-files.htm>). We obtained this data on May 31, 2024.

5) Corporate political contributions

- a. We downloaded corporate political contributions data from the Database on Ideology, Money in Politics and Elections (DIME). We downloaded the data on September 30, 2021.
 - 6) Patent citations data
 - a. We obtained this data from the USPTO's PatentsView.org (<https://patentsview.org/download/data-download-tables>) on January 27, 2023.
 - 7) Patent similarity
 - a. We obtain data on patent textual similarity from Whalen, Lungeanu, DeChurch, and Contractor (2020) (<https://zenodo.org/record/3552078>) on July 14, 2023.
 - 8) Patent value data from Kogan et al. (2017)
 - a. We obtain patent value data from the Kogan et al. (2017) data repository on GitHub (<https://github.com/KPSS2017/Technological-Innovation-Resource-Allocation-and-Growth-Extended-Data>). We obtained the data on May 30, 2024.
 - 9) Patent originality from Trajtenberg et al. (1997)
 - a. We calculate patent originality based on the measure described in Trajtenberg et al. (1997) using backward and forward citations. We obtain patent citations to other patents from PatentsView (<https://patentsview.org/download/data-download-tables>).
 - 10) Firms' financial information
 - a. We downloaded the data from Compustat Fundamentals Annual through WRDS on May 27, 2021
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.*

The data we use from GoodJobsFirst is displayed and searchable on their website. Downloading the data requires a subscription. When we obtained the data, it was publicly available and free to download and use. We have explicit permission to use the data. Moreover, GoodJobsFirst does not restrict publication of results, has given us full control of the relevant data and does not require approval prior to public release.

Phil Mattera (pmattera@goodjobsfirst.org) oversees the Megadeals database and can attest to our access.

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.*

We describe our data collection process in Section 3.1, detailed variable definitions in Section 3.2 and Appendix A, and the sample selection process in Table 1 of our manuscript.

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.*

Apart from hand-collecting information on “Megadeal” subsidies, we manipulate all data via computer programs and we provide the codes for these manipulations in our response to item 6 below.

Upon Acceptance

To be provided upon acceptance of the paper and prior to publication:

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 provide a detailed replication package that provides instructions (README.pdf) and codes to help replicate the data construction and empirical analyses. In addition, we provide the identifiers (COHORT-GVKEY-CFIPS-FYEAR) for the final sample in the replication package.

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 a comprehensive log file that shows the execution of the entire code in the replication package.

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.*

The data and programs will be maintained by Yoojin Lee, Shaphan Ng, and Aruhn Venkat for at least six years, consistent with National Science Foundation guidelines.