Data and Program Guide to "The Origins of Firm Heterogeneity: A Production Network Approach"

Andrew B. Bernard, Emmanuel Dhyne, Glenn Magerman, Kalina Manova and Andreas Moxnes.

November 10, 2021

1 Overview

- This file contains the necessary information to replicate the empirical results of the paper.
- The empirical analysis uses confidential firm-level and firm-to-firm-level data, which we have acquired through a confidentiality agreement, so we are unable to disclose these datasets. These data are administered by the National Bank of Belgium. For replication purposes, researchers can file an application through the Research Department at the NBB (emmanuel.dhyne@nbb.be). The original data are available for replication on secured servers at the National Bank.
- A detailed description of the data sources and construction is given in Section 2 and Appendix A of the paper.
- We provide the full coding pipeline for the results (graphs, tables) in the paper. We also provide code to generate random data with the same variables as present in the real data. This allows for all the code to run properly, although the results will be based on random draws from distributions instead of the real data.
- Codes are organized into tasks.[1] Each task performs one action on the data, with an input, function and output. Within a task, variants of that action are also performed (e.g. main vs robustness). A series of tasks generates the coding and data pipeline for the project.
- Results have been obtained using Stata 16 and Matlab. The program installs the necessary ado-files to run the codes in Stata. Matlab requires the Optimization toolbox.

2 Description per task

- Copy the tasks folder to some location on your machine.
- **main.do:** change the absolute path under "project folder" to this location. This do-file then executes all tasks of the project in sequential order.

[1] For a task-based approach to coding and project pipelines, see for instance this description, or this great talk by Patrick Ball.

2.1 task0_randomdata

- Generate random data with the same variable names as the real data. While the distributions of the variables are roughly comparable to the real data, there is no correlation across variables nor over time. This random data is for code debugging and replication purposes.

- **1. createdata.do:** creates 4 datasets along the same dimensions as the confidential data at the NBB. We have initialized these with 50,000 firms and 500,000 links across these firms in a given year with a given seed. This ensures there are sufficient observations in each decomposition (some decompositions are quite detailed, e.g. NACE 4-digit by NUTS3).

- **2. copy_to_task1.do:** copies task0/output to task1/input in the absence of real data.

2.2 task1_getdata

- This code loads the datasets used in the paper. The confidential data in /input are random, zip_nuts_Belgium.dta contains real information on zip codes and their location. We also provide public information on the confidential datasets in /docs.

- **1. getdata.do:** load the 4 datasets, extract necessary variables.
  - For annual accounts, all flow variables are annualized from fiscal to calendar years.
  - NACE codes have been concorded to the 2008 version to cope with changes in the NACE classification over time.

- **2. clean.do:** fill in panel gaps if missing in one year for zip or nace code. Keep firms with at least 1 FTE.

2.3 task2_FE_regression

- This task runs the high-dimensional two-way fixed effects regression, and reports some statistics on it.

- **1. FE_regression.do:** estimates $\ln m_{ij} = \ln \psi_i + \ln \theta_j + \ln \omega_{ij}$. The reghdfe procedure iteratively drops unidentified fixed effects (called singletons by the progam), and ultimately retains the giang connected component of the network (also often called the mobility group in applications with employer-employee data).

- **2. table1_var_2wayFE.do:** creates Table 1.

2.4 task3_decomposition

- This task performs the variance decompositionof sales in Section 3 and the appendix.

- **1. datasets.do:** from output of FE regression to format for variance decomposition, and add firm observables needed for the decomposition.

- **2. create_components.do:** create variables for decomposition, and construct the components.
- **3. table2_decomposition.do:** creates Table 2, and also results for other years (Table 15 in appendix).
- **4. table3_correlations.do:** creates Table 3.
- **5. E1 decomp_bysector.do:** creates Tables 13 and 14 in the appendix.
- **6. E3 decomp_differences.do:** creates Table 16 in the appendix.
- **7. E4 decomp_nace_nuts3.do:** creates Table 17 (row 1) in the appendix.
- **8. E4 decomp_nacexnuts3.do:** creates Table 17 (row 2) in the appendix.
- **9. E4 nonparametric.do:** creates Figure 11 in the appendix.
- **10. D2 sellerFE_tau.do:** confirms the results in appendix D2.

2.5 task4_stylized_facts

- This task reproduces the stylized facts in Section 2 and additional results in the appendix. This task uses the decomposition sample of Section 3.
- **1. datasets.do:** start from the output of the 2-way FE regression, and add characteristics on both the supplier and customer side.
- **2. fact1_dispersion.do:** creates Figure 1.
- **3. fact2_correlations.do:** creates Figures 2 and 3.
- **4. fact3_assortativity.do:** creates Figure 4.
- **5. B1 distributions_lnS_lnns_lnns.do:** creates Tables 7, 8, 9, 10 in the appendix.
- **6. B2 fact2_sector_pair_demeaning.do:** recreates the result in appendix B2.
- **7. B3 fringe_buyers.do:** recreates Table 11 in the appendix.
- **8. C exo mobility.do:** recreates Figure 9, Figure 10, and Table 12 in the Appendix.

2.6 task5_SMM

- This task sets up the SMM of Section 5. The SMM itself is estimated in Matlab (see below).
- **1. table4_param_calibrations.do:** this calculates the model hyper parameters from the micro data.
- **2. prep_datasets.do:** creates the datasets for the SMM analysis. We prepare datasets for downstream analysis and upstream analysis separately. The focal unit is the firm for which we have the decomposition.

- **3. SMM_moments.do:** calculates the moments of interest for the SMM.
- **4. bootstrap.do:** creates 1000 bootstrapped samples (with replacement) from the estimation sample, to create standard errors for the SMM estimates.

2.7 task6_SMMresults

- **matching_2019.m**: Main file Matlab for SMM estimation, model fit and counterfactuals.
- The file produces:
  - Figure 7
  - All results contained in Table 5
  - Data for Figure 5: distributions.csv
  - Data for Figure 6: sim_data_noZ.csv and sim_data_noF.csv
  - Data for Figure 8: cf_data_baseline.csv and cf_data_nocorr.csv
- **Fig6Fig8.do:** Stata do file for producing Figure 6 and 8.
- **iterate_network.m:** performs the fixed point associated with eq (8) in the paper.
- **networkformation4.m:** calculates the equilibrium network.
- **vat_fp1.m** performs the fixed points associated with eqs (6) and (7) in the paper.
- **SMM_network.m** calculates the simulated moments and the objective function.

2.8 task7_VATgroups

- This task performs additional robustness results, grouping firms into groups that might own individual companies.
- **1. data.do:** sources the raw datasets again for manipulation in this task.
- **2. clean.do:** cleans this data.
- **3. VATgroups.do:** identifies groups for both sellers and buyers, drops intra-firm flows, and sums all in/outgoing flows with partners.
- **4. FE_regression.do:** estimates the FE model again.
- **5. create components.do:** create variables for decomposition, and construct the components.
- **6. decomposition.do:** performs the decomposition on the VAT groups.