README
Replication Package for "Patent Screening, Innovation, and Welfare" (*The Review of Economic Studies*, forthcoming)

Mark Schankerman Florian Schuett

October 2021

1 Overview

This replication package contains the data and code required to reproduce all the tables, figures, and other numbers reported in the paper (including in online appendices). In addition, it contains instructions on how to access and clean the raw data.

2 Data availability and provenance statements

We collected data from a variety of sources, described below. All the data are publicly available. This section describes how to access the documents and raw data used to construct the datafiles needed to run the programs.

2.1 Data collected from public documents

(1) Data on patent office fees in 2005, 2010, and 2015 were collected from the Office of the Federal Register, Code of Federal Regulations, Title 37: Patents, Trademarks, and Copyright. The source documents can be downloaded from:

2005: https://www.govinfo.gov/content/pkg/CFR-2005-title37-vol1/pdf/CFR-2005-title37-vol1.pdf

2010: https://www.govinfo.gov/content/pkg/CFR-2010-title37-vol1/pdf/CFR-2010-title37-vol1.pdf

2015: https://www.govinfo.gov/content/pkg/CFR-2015-title37-vol1/pdf/CFR-2015-title37-vol1.pdf

(2) Data on the cost of operating the patent program in 2005, 2010, and 2015 were collected from the U.S. Patent and Trademark Office (USPTO). The source documents are the USPTO's Performance and Accountability Reports, which can be downloaded from:

2005: https://www.uspto.gov/sites/default/files/about/stratplan/ar/USPTOFY2005PAR.pdf (p. 65)

2010: https://www.uspto.gov/sites/default/files/about/stratplan/ar/USPTOFY2010PAR.pdf (p. 57)

2015: https://www.uspto.gov/sites/default/files/documents/USPTOFY15PAR.pdf (p. 38)

(3) Data on the average salary of examiners and the number of examiners employed at the USPTO in 2005, 2010, and 2015 were obtained from FederalPay.org, a non-governmental information portal built by federal employees. The information can be obtained by accessing https://www.federalpay.org/employees/occupations/patent-examining/ and selecting the relevant year.

(4) Data on patent grants broken down by industry were collected from the USPTO's report on "U.S. Patenting Trends by NAICS Industry Category." Two sets of data were compiled:

   (a) patent grants by year of grant, total (any ownership category); source documents available at https://www.uspto.gov/web/offices/ac/ido/oeip/taf/naics/naics_own_fgall/naics_own_fg.htm

   (b) patent grants by year of application assigned to U.S. corporations; source documents available at https://www.uspto.gov/web/offices/ac/ido/oeip/taf/naics/naics_own_faall/naics_own_fa.htm

(5) Data on the costs of patent litigation were collected from the 2005, 2011, and 2017 editions of the American Intellectual Property Law Association's "Report of the Economic Survey" (AIPLA, 2005; AIPLA, 2011; AIPLA, 2017).

(6) Data on patent propensity by industry were compiled from Cohen et al. (2000). Based on an extensive survey, Cohen et al. (2000) report the sector-specific percentage of innovations for which firms seek patent protection, distinguishing between process and product innovations. In addition, data on the percentage of R&D time spent on each type of innovation were compiled from Arundel and Kabla (1998).

(7) Data on the percentage of innovations that firms in 12 different industries claim would not have been developed without patent protection were collected from Mansfield (1986).

2.2 Datasets obtained from public sources

(8) Data on value of shipments, inventories, wage bill, intermediate materials, capital stock, relevant deflators, and TFP growth at the 6-digit NAICS level are from the NBER-CES Manufacturing Industry Database 1958-2011 (Becker et al., 2016). The data can be downloaded from https://www.nber.org/nberces/nberces5811/naics5811.csv.

(9) Data on capital rental prices, capital stock, and price deflators at 3-digit NAICS level were obtained from the Bureau of Labor Statistics (BLS). The data can be downloaded from https://www.bls.gov/mfp/special_requests/capbymeasure.xlsx. Table 2.1 has capital rental prices, Table 4.1 has productive capital stock, and Table 8.1 has price deflators.

(10) Data on industry concentration (Herfindahl index and the share of the value of shipments accounted for by the 50 largest firms) were obtained from the Census Bureau. The data can be downloaded from https://www.census.gov/content/dam/Census/programs-surveys/economic-census/data/archived_tables/2007/sector31/2007_31-33_Con_Ratios_US.zip.

(11) Monthly data on capacity utilization at the 3-digit NAICS level were obtained from the Federal Reserve. The data can be downloaded from https://www.federalreserve.gov/releases/g17/ipdisk/utl_sa.txt.

(12) Yearly data on capacity utilization in manufacturing were obtained from the St. Louis Federal Reserve. The data can be downloaded from https://fred.stlouisfed.org/series/CUMFN; click on "download" and select preferred format.

(13) Data on total factor productivity at the 3-digit and 4-digit NAICS level were obtained from the BLS. The data can be downloaded from https://www.bls.gov/mfp/mfp_by_industry_and_measure.xlsx (4-digit NAICS level) and https://www.bls.gov/mfp/special_requests/klemsmfp.xlsx (3-digit NAICS level).

(14) Data on company and other nonfederally funded R&D in manufacturing at the 3-digit NAICS level from 1999 to 2004 were obtained from the National Science Foundation (NSF). The data can be downloaded from https://wayback.archive-it.org/5902/20160210163552/http://www.nsf.gov/statistics/nsf07314/tables/tab12.xls (1999-2003), https://wayback.archive-it.org/5902/20160210163454/

http://www.nsf.gov/statistics/nsf09301/tables/tab10.xls (2004), and https://wayback.archive-it.org/5902/20160211030857/http://www.nsf.gov/statistics/nsf09301/tables/tab25.xls (1953-2004 R&D spending in current and constant 2000 dollars).

(15) Information on the correspondence between industry categories and technology classes was obtained from the USPTO. The download links are:

   (a) Correspondence between NAICS code and sequence number (one to one): https://www.uspto.gov/web/offices/ac/ido/oeip/taf/data/naics_conc/2014/read_me.txt

   (b) Correspondence between sequence number and technology class (one to many): https://www.uspto.gov/web/offices/ac/ido/oeip/taf/data/naics_conc/2014/naics_co14.csv

(16) Information on the correspondence between technology classes and technology fields was obtained from the National Bureau of Economic Research. They can be downloaded from https://data.nber.org/patents/class_match.txt.

(17) Data on the real interest rate were obtained from the World Bank. They can be downloaded from https://data.worldbank.org/indicator/FR.INR.RINR?locations=US; select preferred format under "Download."

(18) Detailed patent data covering the period from 1976 to 2006 were obtained from the NBER Patent Data Project. They can be downloaded from http://www.nber.org/~jbessen/pat76_06_assg.dta.zip.

(19) Statistics on patent applications by industry in 2011 were obtained from the NSF. The data can be downloaded from https://www.nsf.gov/statistics/2015/nsf15307/tables/tab38.xlsx.

(20) Statistics on patent applications from 1963 to 2020 were obtained from the USPTO. The data can be downloaded from https://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm.

(21) Data on GDP deflators were obtained from the St. Louis Federal Reserve. They can be downloaded from https://fred.stlouisfed.org/series/GDPDEF; click on "download" and select preferred format.

(22) Data on royalties and license fees in international transactions in 2008 were obtained from the Bureau of Economic Analysis. They can be downloaded from https://apps.bea.gov/scb/pdf/2010/10%20October/1010_services_tables.pdf. The relevant data are in Table 4.1 on p. 42.

3 Dataset list

This section lists the data files provided as part of the package (in the folder `Data`) and explains what they contain. If applicable, it explains how the files were constructed from the raw data obtained from the sources described above.

`USPTO_fees.xlsx`: Contains the data from (1), complemented with data from (20) used for weighting. The numbers are converted to 2018 dollars using GDP deflator data from (21).

`USPTO_budget.xlsx`: Contains the data from (2) and (3). The numbers are converted to 2018 dollars using GDP deflator data from (21).

`number_of_patents_3and4digits.csv`: Contains the data from (4a).

`fractional count.csv`: Contains the data from (4b).

`litigation_costs(AIPLA_survey_data).xlsx` : Contains the data from (5). The numbers are converted to 2018 dollars using GDP deflator data from (21).

`Patent Propensity.csv`: Contains sector-specific overall patent propensities based on the data from (6). To construct these numbers, we weight the raw patent propensity for process and product innovations from Cohen et al. (2000, Table A1, p. 49) by the percentage of R&D time spent on each type of innovation, as reported in Arundel and Kabla (1998, Table 1, p. 133).

`mansfield.xlsx`: Contains the data from (7), complemented with data from (19).

`naics5811.csv`: Dataset from (8).

`capbymeasure_table2.1.csv`: Table 2.1 from (9).

`capbymeasure_table4.1.csv`: Table 4.1 from (9).

`capbymeasure_table8.1.csv`: Table 8.1 from (9).

`ECN_2007_US_31SR12_with_ann.csv`: Dataset from (10).

`capacity_utilisation_monthly.txt`: Dataset from (11).

`capacity_utilisation.csv`: Dataset from (12).

`mfp_4digits.csv`: Contains the MFP data in the worksheet "Measures by industry" from the 4-digit NAICS level data from (13).

`manual_assign_tfp_estimate.csv`: Contains manually assigned TFP estimates for a selection of 6-digit NAICS codes.[1]

`R&D computations.xlsx`: Contains the three tables of data from (14) in worksheets "1999-2003", "2004", and "Constant 2000 dollars". In addition, it contains the worksheet "Computations," which uses the data in the other worksheets to compute the numbers in the file `R&D data.csv` (see below).

`R&D data.csv`: Contains average R&D spending over the 1999-2004 period by 3-digit NAICS sector, computed in column Q of worksheet "Computations" in `R&D computations.xlsx`.

`pat76_06_assg.dta`: Contains the data from (18).

`N_f.csv`: Contains data on average yearly patent applications by technology field over the 1999-2004 period, computed based on the data from (18). We provide a Stata do file and an excel file to reproduce the numbers in the file from the raw data; see Section 5.

`mapping rule.csv`: File from (15a) containing the correspondence between NAICS codes and the sequence numbers used in the patent count data from (4). Each sequence number corresponds to exactly one NAICS code.

`naics_co14.csv`: File from (15b) containing the correspondence between sequence numbers and technology classes. One sequence number can correspond to multiple technology classes.

`classcat.csv`: File from (16) containing the correspondence between technology classes and technology fields (the variable 'CAT'). Each class corresponds to one technology field but one technology field contains many classes.

`old to new naics.csv`: File used to map the level of NAICS aggregation used by the USPTO to the level we use. Ours is the same as the USPTO aggregation except that sector 325 is separated into just 3254 and others (3251, 3252, 3253, 3255, 3256, and 3259 combined).

[1] Our TFP estimates are residuals of a regression of MFP on growth in capacity utilitization. For cases where this does not yield a positive estimate, we manually assign a value according to the following procedure: if they exist and are positive, use the average of 6-digit estimates for the same 5-digit naics code. Otherwise, use the 3-digit average MFP $(1/20 \times \log(mfp_{2007}/mfp_{1987}))$ from the 3-digit NAICS level data from (13).

`GDPDEF.xls`: Dataset from (21).

`real interest rates.xls`: Dataset from (17).

4 Computational requirements

4.1 Software requirements

- Stata (code was last run with version 14)
- Python 3.7.3
  - `numpy` 1.18.1
  - `scipy` 1.4.1
  - `matplotlib` 3.1.3
  - `pandas` 1.0.3
  - `itertools`
  - `os`
  - `sklearn.linear_model`

  The file "`/programs/requirements.txt`" lists these dependencies; please run "`pip install -r requirements.txt`" as the first step.
- Microsoft Excel 2016

4.2 System and memory requirements

The code was last run on a Dell laptop with an Intel 4-core i7-8650U CPU, 16GB of RAM, and with Microsoft Windows 10 Enterprise Version 10.0.19042. Runtime requirements are provided in Section 5.

5 Description of programs

- Stata:
  - `litigation_cost_parameters.do` is a do file that runs the regression using the data from (5) on which the litigation cost parameters (l_0 and l_1) are based.
  - `Computation of N_f.do` is a do file that counts the number of patents assigned to U.S. corporations, by technology field, in the data from (18).
- Python:
  - `general.py` and `mod1_fcts.py` define the classes and functions on which the code for parameter assignment and construction of calibration targets relies.
  - `parameter_values.py` runs the data analysis and computations that generate (a subset of) the assigned parameters and calibration targets. The output is written to `params_targets.csv`. Runtime: 1 minute.
  - `estimation_and_counterfactuals.py` runs the baseline estimation whose results are in Table 2, Panel A and the counterfactuals which are in Tables 3 and 4. It also runs the external validation tests reported in Section 4.4. All of the output is written to `baseline_and_counterfactuals.csv`. Runtime: about 60 minutes.

  - The programs in the `Robustness` folder (`robustness_lr238.py`, `robustness_pakes.py`, and `robustness_frechet.py`) generate the output for the robustness checks in Table 2, Panel B. Their output is written to homonymous csv files. Runtime: each about 2 minutes.
  - The programs in the `Figures` folder (`figure2.py`, `figure3.py`, `figure4.py`, and `figure5.py`) generate the figures (for both main text and online appendix). The code in these programs relies on the program `class_definitions.py`, also in the `Figures` folder, which defines various classes and functions. Runtime: Figure 2 – about 80 minutes, Figures 3-5 – less than 1 minute each. Output files: `W_contours.png` (Figure 2), `growthrates.png` (Figure 3), `profitfuncs.png` (Figure 4), `lambdalow.png` (Figure 5).

- Excel:
  - `Tables.xlsx` uses the datafiles and the output generated by the Python code to produce the tables and other numbers reported in the paper.
  - `Computation of N_f.xlsx` uses the output generated by the Stata do file `Computation of N_f.do` to compute the numbers in `Data/N_f.csv`.

6 Instructions to replicators

Set up the environment as explained above. Store the data files contained in the replication package (described in Section 3) in a subfolder `Data`. Then:

1. Run the Stata do files `litigation_cost_parameters.do` and `Computation of N_f.do`.
2. Run the Python program `parameter_values.py`.
3. Run the Python program `estimation_and_counterfactuals.py`.
4. Run the three Python programs in the subfolder `Robustness`.
5. Run the four Python programs in the subfolder `Figures`.

The order of execution does not matter. (Although the first two steps serve as inputs for the last three steps, and step 3 serves as input for step 5, the relevant output has been copied to the program files that use them so that each can be run independently.)

Once all the output files have been generated, open `Tables.xlsx` with Microsoft Excel. In the pop-up window that asks about links, click "Update", then click "Edit links..." Select all the source files in the list, then click "Open sources."

7 List of tables and figures

For all tables and figures, as well as for numbers reported elsewhere in the text, we indicate the program that generates them and the output file to which they are written. In the last column, we indicate in which worksheet of the file `Tables.xlsx` the relevant exhibit is generated (if applicable).

Exhibit	**Program**	**Output file**	`Tables.xlsx` **sheet**
TABLE 1			
Assigned parameters			
a, c	`parameter_values.py`	`params_targets.csv`	
T	Statutory patent life		
r, b	`Tables.xlsx`		Worksheet "Table 1"
l_0, l_1	`litigation_cost_parameters.do`	Regression output: l_0 = constant, l_1 = coefficient on `value at stake`	
ϕ_A, ϕ_P	`USPTO_fees.xlsx`		
Empirical targets			
GR, LR, VR	Set based on external information (see online Appendix G)		
RPI, $TFPI$	`parameter_values.py`	`params_targets.csv`	
B	`USPTO_budget.xlsx`		Worksheet "Table 1"
E	`Tables.xlsx`		
TABLE 2			
Panel A. Baseline: estimates and welfare decomposition	`estimation_and_counterfactuals.py`	`baseline_and_counterfactuals.csv`	Worksheet "Table 2"
Panel B. Robustness tests	`robustness_lr238.py`, `robustness_pakes.py`, `robustness_frechet.py`	`robustness_lr238.csv`, `robustness_pakes.csv`, `robustness_frechet.csv`	
TABLE 3	`estimation_and_counterfactuals.py`	`baseline_and_counterfactuals.csv`	Worksheet "Tables 3 and 4"
TABLE 4	`estimation_and_counterfactuals.py`	`baseline_and_counterfactuals.csv`	Worksheet "Tables 3 and 4"
NUMBERS REPORTED ELSEWHERE			
Section 1: Optimal fees over current fees	`estimation_and_counterfactuals.py`	`baseline_and_counterfactuals.csv`	
Section 1: Share of low-type patents exposed to challenges	`estimation_and_counterfactuals.py`	`baseline_and_counterfactuals.csv`	
Section 4.3: Ratio of 75th to 25th percentile of s	`estimation_and_counterfactuals.py`	`baseline_and_counterfactuals.csv`	
Section 4.3: Share of patents below the median value estimated by Bessen (2008)	`estimation_and_counterfactuals.py`	`baseline_and_counterfactuals.csv`	Worksheet "Numbers reported elsewhere"
Section 4.3: Gross welfare from high-type innovation over total gross welfare	`estimation_and_counterfactuals.py`	`baseline_and_counterfactuals.csv`	
Section 4.4: Aggregate patent propensity (empirical)	`Tables.xlsx`		
Section 4.4: Share of high-type inventions			
Simulated	`estimation_and_counterfactuals.py`	`baseline_and_counterfactuals.csv`	

Continued on next page

Exhibit	Program	Output file	Tables.xlsx sheet
Numbers reported elsewhere (continued)			
Empirical, unweighted average and patent-weighted average	mansfield.xlsx		
Section 4.4: Elasticity of patent applications to fees			
Simulated	estimation_and_counterfactuals.py	baseline_and_counterfactuals.csv	
Empirical	External estimates		
Section 4.4: Elasticity of patent grants to R&D			
Simulated	estimation_and_counterfactuals.py	baseline_and_counterfactuals.csv	Worksheet "Numbers reported elsewhere"
Empirical	External estimates		
Section 4.4: Cost saving from a registration system			
Simulated	estimation_and_counterfactuals.py	baseline_and_counterfactuals.csv	
Empirical, examiner salaries saved			
Empirical, search and examination fees as share of pre-grant fees	USPTO_budget.xlsx		
Section 4.4: Ratio of licensing revenue to R&D			
Simulated	estimation_and_counterfactuals.py	baseline_and_counterfactuals.csv	
Empirical	Tables.xlsx		
Section 5.1: change in low-type applications when fees are frontloaded			
Section 5.1: social value of the patent system under a registration system			
Section 5.1: sum of pre-grant and post-grant fees			
Section 5.1: share of high types among applicants under the optimal patent policy			
Section 5.1: examination cost per application under optimal policy over examination cost per application in baseline	estimation_and_counterfactuals.py	baseline_and_counterfactuals.csv	Worksheet "Numbers reported elsewhere"
Section 5.1: change in high-type applications under the optimal patent policy			
Section 5.2: low-type deadweight loss avoided (DA) under PTAB (preponderance standard), as share of total DA			
Section 5.2: low types that always preempt challenges under PTAB (preponderance standard)			
Section 5.2: low-type DA under PTAB (preponderance standard) over low-type DA in baseline			
Figure 1	N/A	N/A	
Figure 2	figure2.py	W_contours.png	
Figure 3	figure3.py	growthrates.png	
Figure 4	figure4.py	profitfuncs.png	
Figure 5	figure5.py	lambdalow.png	

References

American Intellectual Property Law Association (2005). *2005 Report of the Economic Survey.* Arlington, VA.

American Intellectual Property Law Association (2011). *2011 Report of the Economic Survey.* Arlington, VA.

American Intellectual Property Law Association (2017). *2017 Report of the Economic Survey.* Arlington, VA.

Arundel, A. and I. Kabla (1998). What percentage of innovations are patented? Empirical estimates for European firms. *Research Policy 27*(2), 127–141.

Becker, R., W. Gray, and J. Marvakov (2016). NBER-CES Manufacturing Industry Database: Technical notes. National Bureau of Economic Research.

Bessen, J. (2008). The value of U.S. patents by owner and patent characteristics. *Research Policy 37*(5), 932–945.

Cohen, W. M., R. R. Nelson, and J. P. Walsh (2000). Protecting their intellectual assets: Appropriability conditions and why U.S. manufacturing firms patent (or not). NBER Working Paper No. 7552.

Mansfield, E. (1986). Patents and innovation: An empirical study. *Management Science 32*(2), 173–181.