STEPS FOR REPLICATION
KNOWLEDGE SPILLOVERS AND PATENT CITATIONS: TRENDS IN GEOGRAPHIC LOCALIZATION, 1976-2015

This file is to explain the procedure of running codes to attain the result in the paper. Read section 1 for overall description for the analysis in the paper. You need Python and stata program to run the files. For each code file, we included more explanation for each step. Also, definition for variables are included in section 2.

1. Steps for Replication

1.1. **Data extraction and cleaning.** First, we downloaded and cleaned the patent bulk data. For the details of cleaning process, read cleaning.do file. The final cleaned csv files are as below.

- basic.csv
- citing_cited.csv
- citing_cited_cleaned.csv
- class.csv
- invent_location.csv

1.2. **Start with originating and citing patents.**

1.2.1. *Originating patents.*

(1) at least one US inventor - using the location data before mapping CMSA
(2) corporation assignee
(3) granted year in cohort sets (1976,1986,1996,2006)

cohort	the number of patents
1976	44,016
1986	38,160
1996	61,581
2006	80,495

Table 1. Originating patent counts

1.2.2. *Citing patents.*

(1) cited originating patents calculated above
(2) applied in 10-year window
(cf. For 2006 cohort, due to the data problem, citing patents granted up to May 2015 are included)

cohort	the number of patents
1976	149,843
1986	253,989
1996	1,008,675
2006	810,919

TABLE 2. Citing patent counts

1.3. **Construct candidates for finding controls.**

- Resulting file: citing_cited_cleaned.csv
- From the raw data on patent citations, we restrict the pairs according to the following conditions.
  (1) cited patent has at least one US inventor, and assigned to corporations
  (2) citing patent cites patents in (1)
  (3) citing patent has corporation assignee
  (4) excluding self-citations
  (5) citing and cited patent have granted and applied date information

1.4. **Find control patents.**

- Run Finding_Controls.ipynb using Python.
- Data required (import files)
  (1) basic.csv
  (2) class.csv
  (3) citing_cited_cleaned.csv
  (4) citing_cited.csv
- Export files: controls#_a_b.csv
  - # = criteria (1: 3-digit / 2: Any / 3: Primary / 4: Common)
  - a = cohort (1976,1986,1996,2006)
    b = end year (1985,1995,2005,2015)

1.5. **Replication Table 1, 2, 4, and 5.** Run Table#.ipynb to get each table. For each file, we listed the required files.

1.6. **Replication Table 3.** Run Table3_4.do to get Table 3 results and to make 'origin_10per_rand.csv' file, which is an input file for Table4.ipynb. The required files are listed in the do file.

2. ReadMe for each data file

■ basic.csv
variables: wku, apd, isd (observations: 5,264,024)
* wku : patent number
* apd : application date
* isd : issue date

■ citing_cited.csv, citing_cited_cleaned.csv
variables: citing, cited (observations: 40,187,180 and 34,267,448)
* citing: patent number of citing patent
* cited: patent number of cited patent
Notes) * Cited patents in two files have at least one US investor and belong to an institution, and their patent numbers are higher than 3,500,000.
* Citing patents in "citing_cited_cleaned.csv" belong to an institution. ("citing_cited.csv" has citing patents belonging to an individual.)
* There are no self-citations in "citing_cited_cleaned.csv" file. ("citing_cited.csv" has self-citations also.)

■ class.csv
variables: wku, adate, type, fullclass, first_class (observations: 20,093,302)
* wku: patent number
* adate: the date of application (days since Jan. 1st, 1960)
* type: OCL if primary class and XCL if additional class
* fullclass: the 9 digit US classification
* first_class: the first 3 digit of US classification

■ invent_location.csv
variables: wku, cnt, sta, cmsaphantom (observations: 4,865,152)
* wku : patent number
* cnt : country assigned to patent
* sta : state assigned to patent
* cmsaphantom : cmsa(and phantom cmsa) assigned to patent

■ controls(controls#_xxx6_xxx5.csv)
variables: cited,citing,citing_control,period
* cited, citing, citing_control : patent number
* period: windows that the control has been found (1, 3, or 6 months)
Notes) in case the controls are missing, citing_control and period are 0

■ NBER_class.csv
variables: wku, class, subcat
Notes) US classification extracted from the raw data are matched with sub-category using the 1-1 matching table. For patents without class numbers due to several reasons

(mainly deleted during the data cleaning procedure) in the extracted dataset, we used NBER data and filled up the missing spots. After, 386 patents were deleted due to data missing.

- states_oder.csv

This file is to ordering states based on their sizes for tables.

- origin_10per_rand.csv

The mostly cited originating patents (top 10 – 12% in terms of the number of citations)