Data Collection and Refinement

Input files:

newsearchname.csv lists all the names to be searched on Google Scholar.

fixedfacultyinfo.csv contains all economists' personal information collected from their department's webpages, including phd graduation year, position, school affiliation and fields of interest. The fields of interest (column f1, f2 and f3 record primary field, secondary field and third field) are determined by a mixture of key work matching of the economists' interest statement and manual checking of the economists' bio and cv.

getsearchname.csv lists the names used in fixedfacultyinfo.csv (first column) and the searched names used in newsearchname.csv (second column). The names used in fixedfacultyinfo.csv (variable name: fixedname) are used as the authors' identifier throughout the data analysis.

droplist.csv lists the authors with unreliable google search results by their fixedname (usually due to too common name, name change during their career and major publication in CS). These authors are dropped from analysis.

Programs (in execution order):

Run (double click) **"Search results downloader.vbs"** to execute search on Google scholar and save the result pages in /webpage.

Run (double click) **"Parse Webpage.vbs"** to parse the webpage into the file parseddata.csv that contains all paper records (each row is a paper) from the saved result pages.

(Since google has changed the format of the search result during summer 2012, the vbs programs cannot be used to repeat the data collection now. "Parse Webpage.vbs" can always apply on webpages saved prior to summer 2012 though.)

We then deleted the unwanted columns in parseddata.csv (these are fragments from some extra complex urls that are supposed to stay in the link column)

Run (in stata) **"refine data.do"** to read in the input files and parseddata.csv, to make some changes (such as generating new variables, change formats and drop authors in droplist) and to save them in stata formats, facultylist.dta and parseddata.dta

Run (in stata) **"index generation.do"** to read in facultylist.dta and parseddata.dta and generate the csv files that will be used in MLE and FGLS analysis : index.csv and authdata.csv.

Run (in stata) **"summary stat.do"** to generate the summary statistics table (Table 1 in the paper) Current output is copied at the end of the do file.

Output files (they are sufficient for all subsequent analysis):

index.csv contains columns as h-indexes of varying combination of parameters. The list of h-indexes and column names are stored in "index directory.xlsx"

authdata.csv contains author level information in the following order: nrc ranking of the school, tenure dummy, years of working after PhD as in 2011-2012, field weights from behavioral to macro.

These two files contains all the data used in the MLE and FGLS analysis.

Intermediate files:

/webpages contains the search result pages from Google scholar saved around April 26 2012.

parseddata.csv contains paper level information directly extracted from search result webpages, including for each paper written by each author in newsearchname.csv, the paper's number of authors, number of citations and publication year.

parseddata.dta contains paper level information in stata format covering the same set of authors in facultylist.dta

facultylist.dta contains author level information, including each author's fixedname, phd graduation year, school affiliation, position, interest weights assigned to each field.

How to update or add new authors

To add new authors, one should

(1) add the author in fixedfacultyinfo.csv
(2) add the author's citation record in parseddata.csv manually
(3) run "refine data.do" and "index generation.do" to get the updated output files

To update the dataset, one has to revise the vbs programs to accommodate the up-to-date google scholar search result format so that searches can be executed similarly and data can be extracted similarly, which might take a couple days to allow necessary testing. The actual time used to complete the searches and parse the webpages is just a few hours.