Blueprint for dataset construction

1 Download data from Center for Responsive Politics (CRP)

The lobbying and contributions data can be downloaded from https://www.opensecrets.org/myos/. A user guide for the CRP data can be found at https://www.opensecrets.org/resources/datadictionary/UserGuide.doc.

The lobbying data will come in a single zip file covering all years and contain the following text files:

- lob_lobbying.txt
- lob_issue_NoSpecficIssue.txt
- lob_agency.txt

The main lobbying file is lob_lobbying.text but this doesn't contain information regarding the issues or agencies lobbied on a given report. This information is in, respectively, the lob_issue_NoSepcificIssue.txt and lob_agency.txt files.

The contributions data will have the following files for each two-year Congressional cycle with the last two digits of the election year given by YY:

- candsYY.txt
- cmtesYY.txt
- pacsYY.txt

The main contributions file is pacsYY.txt but this only refers to the PAC donor and a representative recipient by ID numbers. The names of the PAC and the representative can be linked using the ID numbers to, respectively, cmtesYY.txt and candsYY.txt.

The following now describes the actual construction of the dataset and can be done using STATA do files.

2 Construction of lobbying dataset

2.1 masterLobbying dataset

One can merge the three lobbying text files described above to create a masterLobbying dataset where an observation is a report-issue. There are a few things that should be noted here:

- The lob_lobbying.txt file essentially has two delimiters "|" and "," (at least from the viewpoint of Excel or STATA). Moreover, "," sometimes appears in the name of the PAC. This can create problems when importing the raw data. These problems can be fixed by using find and replace in a text editor (e.g. notepad) by replacing ",," with ",| |," and doing this twice.

  - You can check everything is OK by importing into Excel and telling Excel that the only delimiter is "|". Everything should now be fine in Excel (importantly, the name fields haven't been split) except you will have a bunch of columns that are filled with ",".

- There is a variable called "Ind" in the lob_lobbying.txt file. Loosely, this variable equals "y" if it is not some type of duplicate report. So one should drop all observations with Ind = n.

- Given each lobbying report can contain multiple issues lobbied and multiple agencies lobbied, one needs to allocate the report's lobbying expenditures across issues and agencies. I allocate the expenditure on a report equally across all issues and agencies and only keep expenditures related to lobbying the agency with AgencyID = 2 which is the agency ID for the US House of Representatives.

- To merge the raw lobbying data (which is recorded on an annual basis) with the raw PAC contributions data (which is recorded on a two-year Congressional cycle basis), I aggregate the lobbying data so that it is at the 2-year Congressional cycle frequency.

2.2 Composition of lobbying expenditures across issues

One can now use the masterLobbying dataset to compute lobbying expenditures on issue k by interest group g in Congressional cycle t which are the L_{kgt} in equation (1) from Section 3 of Lake [2015]. One can also compute the share of lobbying on each issue for each interest group and Congressional cycle which are the $l_{kgt} = \frac{L_{kgt}}{\sum_k L_{kgt}}$ from Section 3 of Lake [2015].

One thing to note here:

- The interest group has both a "client" name and an "UltOrg" name. The client name is the interest group that actually did the lobbying and the UltOrg is the parent organization of the client. For example, the American Bankers Association may be the parent and the clients could be the California Bankers Association, the New York Bankers Association etc. The moneyUltOrg.dta dataset uses the UltOrg as the interest group (which is most often the same as the client). The moneyClient.dta uses the client as the interest group.

3 Construction of contributions dataset

3.1 masterContributions dataset

One can merge the three contributions text files described above to create a masterContributions dataset. Observations in this dataset are at the "transaction level": an observation is a contribution by an interest group to a representative in a given Congressional cycle and so there could be multiple contributions by the same interest group to the same representative in the same Congressional cycle.

There are a few things that should be noted here:

- Issues with importing raw data from CRP text files

  - cmtesYY.txt. These files have the same problem as the lob_lobbying.txt file above. It can be fixed in the same way.

  - pacsYY.txt. Again, the presence of "|" and "," as delimiters can create problems when importing the raw data (but there is no issue with "," occurring within variable fields anymore). This can be fixed as follows. First, use find and replace in a text editor (e.g. notepad) and replacing "„" with ",| |," and do this twice. Second, import into Excel specifying two delimiters, "|" and ",". Third, scan for any remaining errors by adding a beginning row to the Excel file and using the "filter" command to check for blanks in each column.
  - candsYY.txt. No issues.

- The cmtesYY.txt files can give missing values for the UltOrg (i.e. parent) of the contributing PAC. In these cases, per the CRP user guide, the UltOrg is the same as the contributing PAC.

- The cmtesYY.txt files have a variable called "RecipCode" which characterizes the type of PAC. For example, two of these codes are "B" for business PAC and "L" for labor PAC. These variables can be used later to get issue-representative specific measures of contributions and lobbying based on the contribution and lobbying activities of i) business PACs only or ii) labor PACs only.

- The candsYY.txt files have a variable that should indicate whether the representative won their Congressional House race. The online dataset is based on representatives who sit in the House of Representatives, so one needs to make sure that the 435 elected representatives are indeed in the dataset for each session of Congress. However, I have found mistakes in the CRP dataset so one needs to check that each Congressional district has a single winning representative.

3.2 Composition of contributions across representatives

One can now use the masterContributions dataset to compute the contributions given from an interest group g to a representative i in each Congressional cycle t which are the C_{igt} from Section 3 of Lake [2015]. One can also compute the share of an interest group's contributions going to each representative in a given Congressional cycle which are the $c_{igt} = \frac{C_{igt}}{\sum_i C_{igt}}$ from Section 3 of Lake [2015].

One thing to note here:

- The cmtesYY.txt files have a variable called "party" which takes on a missing value if the PAC is a party-related PAC (specifically, party-related PACs are party, leadership, joint fundraising or candidate committees). I only consider contributions by non party-related PACs.

- The pacsYY.txt files have a variable called "DI" which records whether the contribution is a direct or indirect contribution. I drop all indirect contributions.

4 Issue-representative specific measures of contributions and lobbying

One can now compute the contributions received by representative i for issue k in Congressional cycle t as $C_{ikt} = \sum_g l_{kgt} C_{igt}$ which is equation (2) from Section 3 of Lake [2015]. Additionally, one can now compute the lobbying targeted at representative i for issue k in Congressional cycle t as $L_{ikt} = \sum_g c_{igt} L_{gkt}$ which is equation (3) from Section 3 of Lake [2015].

By restricting the summation over interest groups to business PACs or labor PACs one obtains measures of C_{ikt} and L_{ikt} based solely on the contribution and lobbying activities of business or labor PACs.

References

J. Lake. Revisiting the link between PAC contributions and lobbying expenditures. *European Journal of Political Economy*, 37:86–101, 2015.