Codebook for the Database on Ideology, Money in Politics, and Elections (DIME) (Version 2.0)

August 16, 2016

Description:
The Database on Ideology, Money in Politics, and Elections (DIME) is intended as a general resource for the study of campaign finance and ideology in American politics. The database was developed as part of the project on Ideology in the Political Marketplace, which is an on-going effort to perform a comprehensive ideological mapping of political elites, interest groups, and donors. Constructing the database required a large-scale effort to compile, clean, and process data on contribution records, candidate characteristics, and election outcomes from various sources. The resulting database contains over 130 million political contributions made by individuals and organizations to local, state, and federal elections spanning a period from 1979 to 2014. A corresponding database of candidates and committees provides additional information on state and federal elections.

Principal Investigator:
Adam Bonica

Contact:
Email: bonica@stanford.edu
Web-page: http://www.stanford.edu/~bonica/

How to Cite this Dataset:
Bonica, Adam. 2016. *Database on Ideology, Money in Politics, and Elections*: Public version 2.0 [Computer file]. Stanford, CA: Stanford University Libraries. <http://data.stanford.edu/dime>.

For CFscore measures, please cite:
Bonica, Adam. 2014. "Mapping the Ideological Marketplace." *American Journal of Political Science*, 58 (2): 367- 387. (http://onlinelibrary.wiley.com/doi/10.1111/ajps.12062/abstract)

For DW-DIME measures, please cite:
Bonica, Adam. 2016. "Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning" *Working Paper.* (Available at SSRN: http://ssrn.com/abstract=2732913)

Contents

1 Introduction **3**

2 Release Notes **4**

3 List of Data Files **5**

4 Data Sources **6**

5 Variable Codings **7**
5.1 Candidate/Recipient Database 7
5.2 Contributor Database 11
5.3 Contribution Database 12

6 Candidate-Contributor Contingency Matrix **15**

7 Notes on Data Usage **16**

8 Tips for Working with Large Data Files **16**

9 Seat Labels **17**

10 Transaction Codes **18**

11 Reverse Compatibility With Earlier Releases **20**

12 Acknowledgments **20**

1 Introduction

A core objective in constructing the database was to make data on campaign finance and elections (1) more centralized and accessible, (2) easier to work with, and (3) more versatile in terms of the types of questions that can be addressed. To these ends, I have put a great deal of effort into compiling, processing, and augmenting the database. In making the database public, I hope to provide a valuable resource to fellow researchers. A list of the main value-added features of the database is below:

Data processing: Names, addresses, and occupational and employer titles have been cleaned and standardized.

Unique identifiers: Entity resolution techniques were used to assign unique identifiers for all individual and institutional donors included in the database. The contributor IDs make it possible to track giving by individuals across election cycles and levels of government.

Geocoding: Each record has been geocoded and overlaid onto congressional districts. The geocoding scheme relies on the contributor IDs to assign a complete set of consistent geo-coordinates to donors that report their full address in some records but not in others. This is accomplished by combining information on self-reported address across records. The geocoding scheme further takes into account donors with multiple addresses. Geocoding was performed using the Data Science Toolkit maintained by Pete Warden and hosted at http://www.datasciencetoolkit.org/.

Ideological measures: The common-space CFscores allow for direct distance comparisons of the ideal points of a wide range of political actors from state and federal politics spanning a 35 year period. In total, the database includes ideal point estimates for 70,871 candidates and 12,271 political committees as recipients and 14.7 million individuals and 1.7 million organizations as donors.

Corresponding data on candidates, committees, and elections: The recipient database includes information on voting records, fundraising statistics, election outcomes, gender, and other candidate characteristics. All candidates are assigned unique identifiers that make it possible to track candidates if they campaign for different offices. The recipient IDs can also be used to match against the database of contribution records. The database also includes entries for PACs, super PACs, party committees, leadership PACs, 527s, state ballot campaigns, and other committees that engage in fundraising activities.

Identifying sets of important political actors: Contribution records have been matched onto other publicly available databases of important political actors. Examples include:

- Fortune 500 directors and CEOs (Data) (Paper)
- State supreme court justices (Data) (Paper)
- Federal court judges (Data) (Paper)
- Executives appointees to federal agencies (Data) (Paper).
- Medical professionals (Data) (Paper)

Each of the above are available for download. (Note that some of these curated datasets have been compiled using version 1.0 of the DIME.)

Extensions: The DIME+ data repository on congressional activity extends DIME to cover detailed data on legislative voting, lawmaking, and political rhetoric. (See http://data.stanford.edu/dime-plus for details.)

2 Release Notes

- Updated FEC data for federal elections through 2014.
- Updated state records for California and Texas through 2014. Data all other states covers up through the 2012 election cycle.
- NIMSP data for California and Texas replaced with bulk data from CalAccess and the Texas SoS's website.
- 527 records from IRS updated through 2014.
- Various improvements made to identity resolution algorithms to enhance accuracy.
- Added DW-DIME scores for candidates. These scores are estimated using a supervised learning model with DW-NOMINATE as the target variable. (See paper referenced above for details.)
- Added candidate estimates from IRT count-model applied to PAC data as described in "Ideology and Interests in the Political Marketplace" Bonica (2013).
- Added several new columns to the contribDB files containing itemized records. (See variable description below for details.)
- Included new columns in contributor table that list the self-reported name, address, occupation, employer from most recent record in the database.
- Added contingency matrices with raw contribution amounts to `MIMP.RData`.
- Utilized historical shape files for congressional districts from Jeff Lewis for most recent election cycles.
- Added new transaction types for state records.
- DW-NOMINATE scores updated though 114th Congress.

3 List of Data Files

Candidate/Recipient Files: The main recipient file includes cycle-specific entries for all candidates and committees included in the scaling (N=176,830). The raw candidate/recipient file additionally includes cycle-specific entries for candidates and committees that did not meet the requirements for inclusion in the scaling—including labor, corporate, and trade PACs that were excluded from the estimation stage and candidates that did not raise funds from the required number of contributors to be included in the scaling (N=379,761).

- `dime_recipients_1979_2014.csv` - Candidate/Recipient CFscores (included in scaling) - State, Federal, 1979-2014 data file
- `dime_recipients_all_1979_2014.csv` - Candidate/Recipient CFscores (all recipients) - State, Federal, 1979-2014 data file

Contributor Files: Includes entries for each individual and organization that has made political contributions (N = 16,409,481).

- `dime_contributors_1979_2014.csv` - State, Federal, 1979-2014

Contingency Matrix of Contribution Amounts: Includes an R list object organizes the contribution data into an n by m contingency matrix of contribution amounts. (See section below for details).

- `mimp.rdata` - Sparse matrix of contribution amounts

Contribution Records: Contains the itemized contribution records grouped by election cycle. An additional set of files contains records organized by seat type. These files contain all contribution records associated with (1) gubernatorial, (2) judicial, or (3) presidential candidates.

Itemized contribution records grouped by cycle:

- `ContribDB_1980.csv` (Row count: 464,336)
- `ContribDB_1982.csv` (Row count: 324,237)
- `ContribDB_1984.csv` (Row count: 448,774)
- `ContribDB_1986.csv` (Row count: 501,673)
- `ContribDB_1988.csv` (Row count: 673,674)
- `ContribDB_1990.csv` (Row count: 1,172,759)
- `ContribDB_1992.csv` (Row count: 1,548,865)
- `ContribDB_1994.csv` (Row count: 1,438,719)
- `ContribDB_1996.csv` (Row count: 2,306,066)

- `ContribDB_1998.csv` (Row count: 4,086,038)
- `ContribDB_2000.csv` (Row count: 5,046,741)
- `ContribDB_2002.csv` (Row count: 8,226,291)
- `ContribDB_2004.csv` (Row count: 12,187,227)
- `ContribDB_2006.csv` (Row count: 13,013,868)
- `ContribDB_2008.csv` (Row count: 18,836,167)
- `ContribDB_2010.csv` (Row count: 17,631,194)
- `ContribDB_2012.csv` (Row count: 30,562,233)
- `ContribDB_2014.csv` (Row count: 23,161,961)

Itemized contribution records grouped by seat type:

- `contribDB_governor.csv` (Row count: 6,554,827)
- `contribDB_judicial.csv` (Row count: 1,001,946)
- `contribDB_president.csv` (Row count: 12,353,214)

Note: If you have issues uncompressing any of zip files, you may need to download third-party compression tools. For Windows users, 7-zip (http://www.7-zip.org/) will support opening the larger zip files. For Mac users, Unarchiver (https://itunes.apple.com/us/app/the-unarchiver/id425424353?mt=12) and Keka (http://www.kekaosx.com/en/) are both good options.

4 Data Sources

Federal Elections: Contribution records, candidate and committee filings, and election outcomes for federal elections are from the Federal Election Commission (FEC).

State Elections: Contribution records, candidate and committee filings, and election outcomes for state elections for the 2014 election cycle are provided by the National Institute on Money in State Politics (NIMSP) and the Sunlight Foundation. This data is licensed by NIMSP under the Creative Commons Attribution-Non-commercial-Share Alike 3.0 United States License. (See here for details: http://followthemoney.org/Institute/about_data.phtml.) When using data on state elections, please attribute credit accordingly. As of this release, only records for California and Texas are available for the 2014 election cycle. These records were downloaded in bulk from the respective state campaign finance reporting agencies.

527s: Donation records to 527s are from the Center for Responsive Politics (2002-2010) and the IRS (2011-2014). The Center for Responsive Politics licenses its data under the Creative Commons Attribution-Non-commercial-Share Alike 3.0 United States License. Please attribute credit

accordingly.

New York City Elections: Contribution records for New York City elections were downloaded from the New York City Campaign Finance Board's website (http://www.nyccfb.info/).

Other Data:

DW-NOMINATE scores are provided by Keith Poole and Howard Rosenthal and are available for download at http://www.voteview.com.
Cite: Poole, Keith T., and Howard Rosenthal. 2007. *Ideology & Congress*. 2nd rev. ed. New Brunswick: Transaction Publishers.

Data on historical boundaries for congressional districts are from:
Jeffrey B. Lewis, Brandon DeVine, Lincoln Pitcher, and Kenneth C. Martis. (2013) Digital Boundary Definitions of United States Congressional Districts, 1789-2012. Version 1.2 (March 20, 2015). Retrieved from http://cdmaps.polisci.ucla.edu.

Shape files for census tracts are from Census.gov: (http://www.census.gov/rdo/data/).

5 Variable Codings

5.1 Candidate/Recipient Database

The candidate/recipient database includes information on voting records, fundraising statistics, election outcomes, and other candidate characteristics. All candidates are assigned a unique identifiers that track them as they move across offices. The recipient IDs can also be used to match against the database of contribution records. The database also includes entries for PACs, super PACs, leadership PACs, 527s, state ballot campaign committees, and other recipient committees that engage in fundraising activities. Entries for candidates with two or more active fundraising committees during a given election cycle have been de-duped. The methodology used to estimate common-space CFscores for contributors is described in "Mapping the Ideological Marketplace" (Bonica 2014). The methodology used to estimate the DW-DIME scores is described in "Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning" (Bonica 2016).
`Variable Descriptions:`

1) `election:` Election cycle preceded by two-letter state code. Federal candidates have 'fd' as the state code.

2) `cycle:` Four digit number that indicates the two-year election cycle during which the contribution was recorded.

3) `fecyear:` Year listed by the FEC indicating the year of campaign's the target election. The 'election' variable indicates the election cycle during which the contribution was received. But the election can occur in a future cycle—as is the case for senators that fundraise during their first four years in office.

4) `Cand.ID`: The candidate ID assigned by the FEC.

5) `FEC.ID`: The ID assigned by the FEC to the candidates campaign committee.

6) `NID`: (CRP) Unique candidate IDs assigned by the Center for Responsive Politics.

7) `ICPSR`: Adjusted ICPSR legislator ID. Candidates that have never served in Congress are assigned IDs based off of their FEC IDs. The four-digit election cycle is appended to the end of each ID to distinguish separate entries from the same candidate. Candidates that are active in multiple election cycles will have multiple IDs.

8) `ICPSR2`: Adjusted ICPSR legislator ID. Each candidate receives a unique ID that is constant across election cycles. Following Poole and Rosenthal, party switchers are assigned new ICPSR2 IDs after switching parties.

9) `bonica.rid`: Unique ID assigned to candidates/recipients. Each candidate/recipient receives a unique ID that is constant across election cycles, levels of government, and offices sought. With the exception of party-switchers the bonica.rid values have a one-to-one correspondence with the ICPSR2 scores. (Use this variable to merge with the database of contribution records).

10) `bonica.cid`: The unique contributor ID for the candidate. This variable can be used to match candidates with their personal contributions records. (Note: The construction of this variable is not yet complete. It includes a partial set candidates that could be easily linked with their contribution records using an automated matching scheme. Missing values do not necessarily mean that a candidate has not made contributions).

11) `name`: Name of the candidate/recipient.

12) `lname`: Last name of the candidate/recipient.

13) `ffname`: Concatenates first name, middle name, suffix, and title.

14) `fname`: First name of the candidate/recipient.

15) `mname`: First name of the candidate/recipient.

16) `nname`: Nick name of the candidate/recipient.

17) `title`: Title of the candidate/recipient. (e.g. Mr., Mrs., Dr., Esq).

18) `suffix`: Suffix of the candidate/recipient.

19) `party`: Party of candidate/recipient (100 = Dem, 200 = Rep, 328 = Ind).

20) `state`: Two-letter state abbreviations.

21) `seat`: Office sought. Committees are listed as federal:committee. See section on seat codes below.

22) `district`: District code: two-letter state code followed by congressional district number. District numbers for senate candidates take on the value of 'S' followed by the year of the seat will be up for election.

23) `incum.chall`: Incumbency status. ('I' = Incumbent, 'C' = Challenger, 'O' = Open Seat Candidate, " - not up for election).

24) `recipient.cfscore`: Estimated ideology of candidate/recipient based on donations received.

25) `contributor.cfscore`: Estimated ideology of candidate/recipient based on personal donations given to other candidates/recipients.

26) `recipient.cfscores.dyn`: Period-specific estimates of ideology. (Candidate/recipient scores are re-estimated in each election cycle while holding contributor scores constant.)

27) `dwnom1`: (PR) First dimension common-space DW-NOMINATE score. (Based on joint scaling of the 1st to the 112th Congresses.)

28) `dwnom2`: (PR) Second dimension common-space DW-NOMINATE score.

29) `ps.dwnom1`: (PR) First dimension Nokken-Poole period-specific DW-NOMINATE score. (Scores for the House and Senate are scaled separately and thus should not be directly compared.)

30) `ps.dwnom2`: (PR) Second dimension Nokken-Poole period-specific DW-NOMINATE score.

31) `dwdime`: DW-DIME scores for recipients. These scores are described in detail in "Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning" Bonica (2016).

32) `irt.cfscores`: Estimates of ideology for recipients/candidates from IRT count-model applied to PAC data. (See "Ideology and Interests in the Political Marketplace" Bonica (2013) for details.)

33) `num.givers`: Number of distinct donors that gave to the candidate during a specific election cycle.

34) `num.givers.total`: Number of distinct donors that gave to the candidate/recipient aggregating over the candidate/recipient's career.

35) `n.data.points.personal.donations`: Number of personal contributions records made by candidate.

36) `n.data.points.personal.donations.unq`: Number of distinct recipients to whom the candidate personally donated.

37) `cand.gender`: Candidate gender codings. (With the exceptions of candidates that have served in Congress, all gender codes are based on an automated coding scheme that incorporates information gender ratios of first names as reported by the U.S. Census and gender-specific titles (e.g. Mrs., Mr., Jr., Sr.) reported in the contribution records.

38) `total.disbursements`: Total campaign disbursements (in dollars) for the given election cycle.

39) `total.pc.contribs`: Total receipts from party committees.

40) `contribs.from.candidate`: total receipts from candidate contributions.

41) `unitemized`: total unitemized receipts.

42) `non.party.ind.exp.for`: non-party independent expenditures made in support of the candidate.

43) `non.party.ind.exp.against`: non-party independent expenditures made against the candidate.

44) `ind.exp.for`: total independent expenditures made to support the candidate.

45) `ind.exp.against`: total independent expenditures made against the candidate.

46) `comm.cost.for`: total communication costs made on behalf of the candidate.

47) `comm.cost.against`: total communication costs made to oppose the candidate.

48) `party.coord.exp`: total party coordinated expenditures.

49) `party.ind.exp.against`: total independent expenditures made by opposing party against the candidate.

50) `total.receipts`: total dollars raised by candidate during an election cycle.

51) `total.indiv.contributions`: total individual receipts.

52) `total.pac.contributions`: total PAC receipts.

53) `ran.primary`: indicator variable for whether the candidate was active in primary elections.

54) `ran.general`: indicator variable for whether the candidate was active in general elections.

55) `p.elec.stat`: FEC primary election code (W = Win) (L = Lose).

56) `s.elec.stat`: FEC special election code (W = Win) (L = Lose).

57) `r.elec.stat`: FEC run-off election code (W = Win) (L = Lose).

58) `gen.elec.stat`: FEC general election code (W = Win) (L = Lose).

59) `gen.elect.pct`: FEC reported vote share in general election.

60) `winner`: 'W' = won election; 'L' = lost election.

61) `district.partisanship`: Kernell's (2009) measure of district partisanship for the current election cycle. (Interpolated using district presidential vote shares for 2010-2014.)

62) `district.pres.vs`: district-level percentage of the two-party vote share won by the Democratic presidential nominee in the most recent presidential election.

63) `CandStatus`: indicates the status of the candidate's campaign assigned by the FEC. ('C' = Statutory candidate; 'F' = Statutory candidate for future election; 'N' = Not yet a statutory candidate; 'P' = Statutory candidate in prior cycle).

64) `recipient.type:` (cands = candidate, comm = committee).

65) `igcat:` FEC Interest group category code (C = Corporation, L = Labor organization, M = Membership organization, T = Trade association, V = Cooperative, W = Corporation without capital stock).

66) `comtype:` FEC code for type of committee (FEC description)

67) `nimsp.party:` (nimsp) three-letter party code assigned by the NIMSP.

68) `nimsp.candidate.ICO.code:` (nimsp) incumbency status assigned by the NIMSP.

69) `nimsp.district:` (nimsp) district number assigned by the NIMSP.

70) `nimsp.office:` (nimsp) state-office sought.

71) `nimsp.candidate.status:` (nimsp) election outcome.

72) `before.switch.ICPSR:` ICPSR ID prior to switching parties (included for party-switchers only).

73) `after.switch.ICPSR:` ICPSR ID after switching parties (included for party-switchers only).

74) `party.orig` Original party before switch.

5.2 Contributor Database

Includes rows for 16,409,481 individuals and organizational donors included in the database. *Note that all donors are assigned ideal point estimates, including one-off donors who contributed only to single candidate or committee.* The *is.projected* column is used to indicate donors that were excluded from the estimation stage and later projected onto the recovered space as supplementary observations. Contributors who have only given to a single recipient are assigned the ideal point of the recipient. Researchers should aware of this when deciding which donors/ideal points to include in their study. The *num.distinct* column indicates the number of observations that were used to estimate the contributors ideal point. Typically, donating to eight or more distinct recipients is sufficient to recover a reliable ideal point estimate.

1) `bonica.cid:` Unique contributor IDs for each donor in the database.

2) `contributor.type:` Contributor type ('I' = individual, 'C' - committee/organization).

3) `num.records:` The number of records in the contribution database by the donor.

4) `num.distinct:` The number of distinct recipients included in the scaling receiving contributions from the donor.

5) `most.recent.contributor.name:` Contributor's self-reported name from most recent record.

6) `most.recent.contributor.address:` Contributor's self-reported street address from most recent record.

7) `most.recent.contributor.city:` Contributor's self-reported name city/municipality on most recent record.

8) `most.recent.contributor.zipcode:` Contributor's self-reported zip-code (5 or 9 digits) from most recent record.

9) `most.recent.contributor.state:` Contributor's self-reported state from most recent record.

10) `most.recent.contributor.occupation:` Contributor's self-reported occupational title from most recent record.

11) `most.recent.contributor.employer:` Contributor's self-reported employer from most recent record.

12) `most.recent.transaction.id:` transaction.id value of the most recent record from the contribution database

13) `contributor.gender:` Contributor gender ('F'=Female, 'M'=Male, 'U'=Uncoded/Unknown)

14) `is.corp:` Indicates whether the contributor is identified as either a corporate entity or q trade organization (only applies to committees). Takes on the value 'corp' for corporations and trade organizations and is blank otherwise. These donors are excluded from the scaling.

15) `contributor.cfscore:` Contributor CFscore.

16) `is.projected:` Indicates whether the was excluded from the estimation stage but was later projected onto the recovered space as supplementary observations. This will take on the value of 1 for PACs and organizations directly affiliated with corporations or trade organizations and individual donors who gave to a single recipient.

17) `first.cycle.active:` The first recorded cycle in which the donor was active.

18) `last.cycle.active:` The last recorded cycle in which the donor was active.

19) `amount.cycle.`

$$cycle$$

: Total amount contributed in a given election cycle.

5.3 Contribution Database

The contribution database includes a complete set of contribution records grouped by election cycle. Each row represents and individual transaction between a donor and recipient.

NOTE: There are several columns of ID numbers for recipients. Use bonica_rid and bonica_cid as the default. The other columns are included so that the database can be linked back to other data sources.

Some of the variables included in the contribution database are provided by the Center for Responsive Politics (http://www.opensecrets.org) and the National Institute for Money in State Politics (http://www.followthemoney.org). These variables are indicated by placing the strings (CRP) or (NIMSP) before the variable name. If any of these variables are used, please attribute credit accordingly.

1) `cycle`: Election Cycle.

2) `transaction.id`: A primary key that contains a unique transaction id for each record.

3) `transaction.type`: FEC code for transaction type. (See section below for details.)

4) `amount`: Dollar amount of the contribution.

5) `date`: Transaction date of the contribution.

6) `bonica.cid`: A unique contributor id assigned to each individual and organization in the database.

7) `contributor.name`: Complete name of contributor (last, first); suffix and title removed.

8) `contributor.lname`: Last name of contributor.

9) `contributor.fname`: First name of contributor.

10) `contributor.mname`: Middle name or initial of contributor.

11) `contributor.suffix`: Suffix of contributor (e.g. Jr., Sr.).

12) `contributor.title`: Title of contributor (e.g. Mr., Mrs., Dr., Esq).

13) `contributor.ffname`: Concatenates first name, middle name, suffix, and title.

14) `contributor.type`: ('I' = individual; 'C' = committee or organization).

15) `contributor.gender`: Contributor gender coding ('M' = male; 'F' = female; 'U' = unknown). Gender codes are based on an automated coding scheme that incorporates information gender ratios of first names as reported by the U.S. Census and gender-specific titles (e.g. Mrs., Mr., Jr., Sr).

16) `contributor.address`: Contributor's self-reported street address.

17) `contributor.city`: Contributor's self-reported name city/municipality.

18) `contributor.state`: Contributor's self-reported state.

19) `contributor.zipcode`: Contributor's self-reported zip-code (5 or 9 digits).

20) `contributor.occupation`: Contributor's self-reported occupational title.

21) `contributor.employer`: Contributor's self-reported employer.

22) `is.corp`: Indicates whether the contribution is made by a corporate entity or q trade organization (only applies to committees). Takes on the value 'corp' for corporations and trade organizations and is blank otherwise.

23) `recipient.name`: Name of the recipient candidate or committee.

24) `bonica.rid`: Unique ID for recipients. Can be matched against candidate database which contains more detailed information on candidates, elections, and constituencies.

25) `recipient.party:` Party of the recipient (100=DEM; 200=REP; 328 = IND). (Match against candidate database for more detailed party codings.)

26) `recipient.type:` ('CAND' = candidate; 'COMM' = PAC, organization, or party committee)

27) `recipient.state:` Two-letter state abbreviation of the recipients.

28) `seat:` Elected office sought by candidate.

29) `election.type:` ('P' = primary elections; 'G' = general elections).

30) `latitude:` Geo-location (latitude).

31) `longitude:` Geo-location (longitude).

32) `gis.confidence:` A measure of confidence of the accuracy of the gis coordinates. (See http://www.datasciencetoolkit.org for details.

33) `contributor.district.90s:` Contributor's geocode mapping onto a congressional district with respect to boundaries for 1992-2000.

34) `contributor.district.00s:` Contributor's geocode mapping onto a congressional district with respect to boundaries for 2002-2010.

35) `contributor.district.10s:` Contributor's geocode mapping onto a congressional district with respect to boundaries for 2012-2020.

36) `censustract:` Contributor's geocode mapping onto a census tract.

37) `efec.memo:` Memo field from FEC electronic filings.

38) `efec.memo2:` Auxiliary memo field from FEC electronic filings.

39) `efec.transaction.id.orig:` Original transaction id from FEC electronic filings.

40) `bk.ref.transaction.id:` Indicates whether the contribution record previously appeared in the database. The value link back to transaction.id. This can be used to remove duplicate entries.

41) `efec.org.orig:` Original recipient name from from FEC electronic filings.

42) `efec.comid.orig:` Original committee ID from FEC electronic filings.

43) `efec.form.type:` Form type from FEC electronic filings.

44) `contributor.cfscore:` Contributor's ideal CFscore.

45) `candidate.cfscore:` Candidate/recipient's CFscore.

6 Candidate-Contributor Contingency Matrix

The file includes an R list object that contains sparse matrix objects that organize the contribution data into n by m contingency matrices of amounts where the rows index contributors, the columns index candidates/recipients, and each entry R_{ij} stores the total amount contributor i gives to recipient j. Note that the cell values do not represent raw dollar amounts. Rather, they are the transformed values used to recover the common-space CFscores. The transformation is based on a normalization scheme that helps to adjust for variation in contribution limits by converting contribution amounts to count values. The conversion is based on federal contribution limits. Contributions between \$1 and \$100 are coded as 1, contributions between \$101 and \$200 are coded as 2, and so on. Contributions of \$5,000 or greater are capped at 50. (See "Mapping the Ideological Marketplace" (Bonica 2014) for details.) Another set of sparse matrix objects report the raw dollar amounts for each cell.
The R list object contains four sparse matrices:

- `mimp$contrib.matrix` - Sparse matrix of count values with columns indexed by `ICPSR`. (Includes separate columns for each candidate/recipient-cycle observation.)
- `mimp$cm` - Sparse matrix of count values with columns indexed by `ICPSR2`. (Collapses columns such that each candidate/recipient has a single column and cell values are aggregate amounts given across cycles.)
- `mimp$contrib.matrix.am` - Sparse matrix of raw dollar amounts with columns indexed by `ICPSR`. (Includes separate columns for each candidate/recipient-cycle observation.)
- `mimp$cm.am` - Sparse matrix of raw dollar amounts with columns indexed by `ICPSR2`. (Collapses columns such that each candidate/recipient has a single column and cell values are aggregate amounts given across cycles.)
- `mimp$cands` - Candidate/recipient database as a data.frame object.
- `mimp$contribs` - Contributor database as a data.frame object.

7 Notes on Data Usage

Although I strive for completeness and accuracy in compiling the database, it is not guaranteed to be 100% comprehensive or without error. The donor and candidate IDs were assigned using a automated entity resolution framework that relies on probabilistic record-linkage algorithms. Researchers interested in collecting contribution records for a specific group of donors are encouraged to hand-check the IDs assigned by the identity resolution algorithm to screen for possible errors. Corrections are welcome.

Duplicate contribution records are also known to occur in the database. This is often due to cross-reporting by one or more agency. For example, contributions to federal-connected 527s will at times be reported by the FEC under the transaction code 19. Contributions given through conduits also appear the FEC data twice with different recipients listed. The current release includes a column named `bk.ref.transaction.id` to aid in removing duplicated records.

Note also that estimated ideal points for candidates and contributors are based on varying amounts of data. For easy reference, the contributor and candidate/recipient files include columns indicating the number of contribution observations that were used to calculate each score. These files also indicate which donors/recipient were included in the scaling. In many cases, a donor may have only contributed to a single recipient and thus was excluded when estimating the ideal point model. Researchers should take this into account when deciding which donors to include in a study and may consider defining a minimum threshold for the number of contributions required for inclusion.

8 Tips for Working with Large Data Files

Due to the size of the contribution files, working with a statistical software package such as R is often necessary. The delimited CSV files will load into most spreadsheet oriented applications but will usually be truncated due to software limitations files with more than 1 million rows. Even when using a statistical software package, some of the contribution files from more recent election cycles can run up against memory limits on machines with less than 16GB of memory. Fortunately, there are several workarounds for accessing the data from disk or using publicly available cloud-based applications. One approach that has worked well is to import the CSV files into a SQLite database. This can be done within R using the RSQLite package. Alternatively, there are several freely available SQLite applications with built-in support for working with large CSV files. Another user-friendly approach is Google Fusion Tables, a cloud-based platform designed for working with large data files. Google provides the service free-of-charge and provides detailed documentation for getting started and taking advantage of its more advanced features.

9 Seat Labels

List of seat labels assigned to candidates with respect to their target office:

Candidates	
federal:house	U.S. House of Representatives
federal:senate	U.S. Senate
federal:president	U.S. President
state:upper	upper chamber of state legislature
state:lower	lower chamber of state legislature
state:judicial	state high courts
state:judicial:lower	state lower courts
state:office	state-wide office (see `nimsp.seat` detailed codes)
state:governor	state governor
nyc:city	local NYC office
state:office:sheriff	sheriff
state:office:da	district attorney
Committees	
federal:committee	federal committee
state:committee	state committee
federal:527	527 organization

10 Transaction Codes

10 NON-FEDERAL RECEIPT FROM PERSONS LEVIN (L-1A)
11 TRIBAL CONTRIBUTION
12 NON-FEDERAL OTHER RECEIPT LEVIN (L-2)
13 INAUGURAL DONATION ACCEPTED
15 CONTRIBUTION
15C CONTRIBUTION FROM CANDIDATE
15E EARMARKED CONTRIBUTION
15F LOANS FORGIVEN BY CANDIDATE
15I EARMARKED INTERMEDIARY IN
15J MEMO (FILER'S \% OF CONTRIBUTION GIVEN TO JOIN
15T EARMARKED INTERMEDIARY TREASURY IN
15Z IN-KIND CONTRIBUTION RECEIVED FROM REGISTERED
16C LOANS RECEIVED FROM THE CANDIDATE
16F LOANS RECEIVED FROM BANKS
16G LOAN FROM INDIVIDUAL
16H LOAN FROM CANDIDATE/COMMITTEE
16J LOAN REPAYMENTS FROM INDIVIDUAL
16K LOAN REPAYMENTS FROM CANDIDATE/COMMITTEE
16L LOAN REPAYMENTS RECEIVED FROM REGISTERED EN
16R LOANS RECEIVED FROM REGISTERED FILERS
16U LOAN RECEIVED FROM UNREGISTERED ENTITY
17R CONTRIBUTION REFUND RECEIVED FROM REGISTERED
17U REF/REB/RET RECEIVED FROM UNREGISTERED ENTITY
17Y REF/REB/RET FROM INDIVIDUAL/CORPORATION
17Z REF/REB/RET FROM CANDIDATE/COMMITTEE
18G TRANSFER IN AFFILIATED
18H HONORARIUM RECEIVED
18J MEMO (FILER'S \% OF CONTRIBUTION GIVEN TO JOIN
18K CONTRIBUTION RECEIVED FROM REGISTERED FILER
18S RECEIPTS FROM SECRETARY OF STATE
18U CONTRIBUTION RECEIVED FROM UNREGISTERED COMMI
19 ELECTIONEERING COMMUNICATION DONATION RECEIVE
19J MEMO (ELECTIONEERING COMMUNICATION \% OF DONAT
20 DISBURSEMENT - EXEMPT FROM LIMITS
20A NON-FEDERAL DISBURSEMENT LEVIN (L-4A) VOTER R
20B NON-FEDERAL DISBURSEMENT LEVIN (L-4B) VOTER I
20C LOAN REPAYMENTS MADE TO CANDIDATE
20D NON-FEDERAL DISBURSEMENT LEVIN (L-4D) GENERIC
20F LOAN REPAYMENTS MADE TO BANKS
20G LOAN REPAYMENTS MADE TO INDIVIDUAL
20R LOAN REPAYMENTS MADE TO REGISTERED FILER
20V NON-FEDERAL DISBURSEMENT LEVIN (L-4C) GET OUT
22G LOAN TO INDIVIDUAL
22H LOAN TO CANDIDATE/COMMITTEE
22J LOAN REPAYMENT TO INDIVIDUAL

22K LOAN REPAYMENT TO CANDIDATE/COMMITTEE
22L LOAN REPAYMENT TO BANK
22R CONTRIBUTION REFUND TO UNREGISTERED ENTITY
22U LOAN REPAID TO UNREGISTERED ENTITY
22X LOAN MADE TO UNREGISTERED ENTITY
22Y CONTRIBUTION REFUND TO INDIVIDUAL
22Z CONTRIBUTION REFUND TO CANDIDATE/COMMITTEE
23Y INAUGURAL DONATION REFUND
24A INDEPENDENT EXPENDITURE AGAINST
24C COORDINATED EXPENDITURE
24E INDEPENDENT EXPENDITURE FOR
24F COMMUNICATION COST FOR CANDIDATE (C7)
24G TRANSFER OUT AFFILIATED
24H HONORARIUM TO CANDIDATE
24I EARMARKED INTERMEDIARY OUT
24K CONTRIBUTION MADE TO NON-AFFILIATED
24N COMMUNICATION COST AGAINST CANDIDATE (C7)
24P CONTRIBUTION MADE TO POSSIBLE CANDIDATE
24R ELECTION RECOUNT DISBURSEMENT
24T EARMARKED INTERMEDIARY TREASURY OUT
24U CONTRIBUTION MADE TO UNREGISTERED
24Z IN-KIND CONTRIBUTION MADE TO REGISTERED FILER
29 ELECTIONEERING COMMUNICATION DISBURSEMENT(S)

NON-FEC CODES
15S CONTRIBUTION TO STATE ELECTIONS (CATCHALL)
15L CONTRIBUTION TO LOCAL ELECTIONS (CATCHALL)
15PD CONTRIBUTION MADE AS PAYROLL DEDUCTION

11 Reverse Compatibility With Earlier Releases

The database has undergone substantial changes and revisions from the original release. Improvements and other modifications were made to the identity resolution algorithms to take advantage of addition information and increased computational resources. The BONICA.CID values for individual records may differ between releases. The TRANSACTION.ID has remained unchanged and can be used as a crosswalk for tracking changes in *bonica.cid* values since the original release of the database.

12 Acknowledgments

I thank the Sunlight Foundation, the National Institute on Money in State Politics, and the Center for Responsive Politics for making their data publicly accessible. I also thank Keith Poole, Howard Rosenthal, Charles Stewart, Jeff Lewis, Jonathan Woon, and Georgia Kernell for providing data. I also acknowledge the generous support I received as a fellow at the Institute for Research in the Social Sciences (IRISS) at Stanford University and the Hoover Institution. Lastly, I thank Ron Nakao for his generous assistance in hosting the database.