Analyzing 2012 FEC data as a graph

Richard K. Belew

12 Nov 2012

1 Introduction

1.1 Motivation

Every election cycle the amount of money spent trying to sway results increases. Because of the Citizens United decision but other factors, too, the volumes of money involved makes news daily. You read all the time about just how much money is sloshing through the system, the big donors providing it, the tricky ways it ultimately filters through to candidates via various PACs and other committees. But I don't want to just know about the activities of the Koch Brothers, Adelson or Karl Rove in isolation, I want to try to understand the flow of vast dollars through the system of political finance; I want to build a graph (in the mathematical sense) for this data. Excited by the data provided as part of the "Follow the money" contest sponsored by Center for Investigative Reporting (CIR) and Investigative Reporters and Editors, Inc. http://www.kaggle.com/c/cir-prospect, this seemed possible. This document reports on the results from a first-pass analysis.

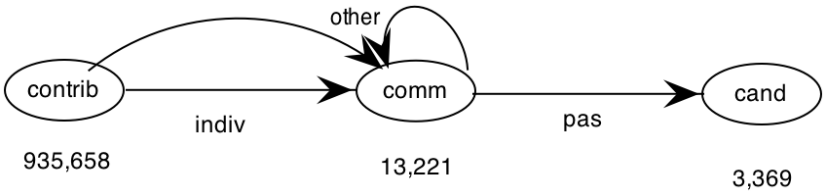

Figure 1: CIR graph, input data

1.2 Data analysis

This analysis is focused on viewing the data as a graph like that shown in Figure 1. Nodes in this graph represent either:

- CANDIDATES,
- COMMITTEES providing direct or indirect contributions to those CANDIDATES, and
- CONTRIBUTORS the ultimate source of contributions to COMMITTEES.

The result is a graph with CONTRIBUTORS acting as a source of funds, passing these funds via INDIV contributions to COMMITTEES, and then these COMMITTEES making contributions to CANDIDATES as recorded in PAS records. OTHER transactions include the PAS records and also transactions of two different types: those from CONTRIBUTORS to COMMITTEES, and intra-COMMITTEE transfers.

The CANDIDATES and COMMITTEES are identified via unique IDs in the provided database. The data in the OTHER contributions appear to be documentation of both COMMITTEE-to-CANDIDATE (PAS) records, together with intra-COMMITTEE passing of funds. Almost a million CONTRIBUTORS were inferred to be those sharing an identical (string equal) name in the INDIV report, contributing to some COMMITTEE. The number of nodes of each type are shown below that node. Differences in these numbers from the mysql rows are detailed below and generally have to do with error checking/data cleansing.

Node counts are shown below the nodes in Figure 1for the number of CONTRIBUTORS, COMMITTEES and CANDIDATES extracted from the original data set. From this set a much smaller more tractable subset was developed. The primary filtering and contributors involved the largest donors being distinguished from all of the rest.

In order to make analysis of this data more tractable, emphasis was placed on watching the most money going through the fewest hands, as it comes from contributors, through committees, to candidates. It is an obvious and widely appreciated fact that a tiny number of individuals give enormous contributions relative to the rest of us. It turns out that 268 CONTRIBUTORS provide one tenth of the $2.3B total campaign contributions. The rest of us can be organized into some that strata each providing other 10% fractions of the total. The analysis below will break out the top 10% BIGGIVERS, and treat the other nine decile strata as super individuals. Note that in this number are many that are functioning as committees. This simply means that they did not have an identified committee ID.

The set of CANDIDATES considered was restricted only by a focus on only 2012 election data. This distinction also proved problematic, and there remain many more candidates than there should be. For example, several candidates are listed in both House and Senate races, sharing a common committee finance committee.

The most difficult selection process involved filtering COMMITTEES to be considered. The heuristic used here was to take the union of three types of committees:

- those COMMITTEES associated as the primary one for a CANDIDATE
- those COMMITTEES receiving >$1M from CONTRIBUTORS
- those COMMITTEES providing >$100K in contributions to CANDIDATES.

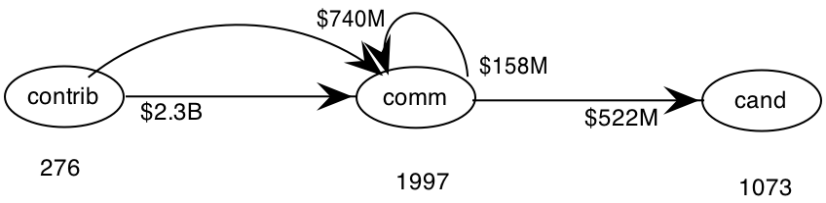

Figure 2: CIR_3364

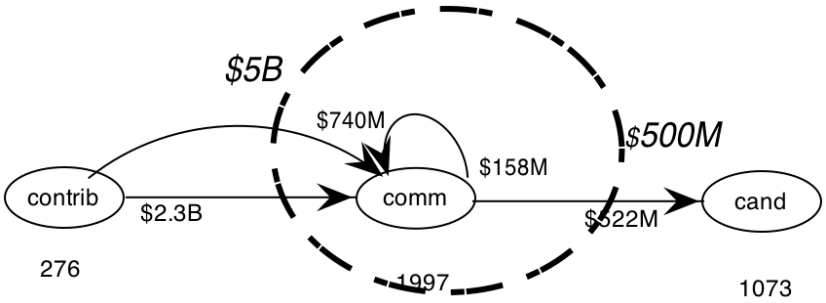

Figure 3: All committees

Distributions associated with these two last sets are shown in details below.

This filtering of the graph results in one we will call CIR_3364, shown in Figure 2.

As shown in Figure 3, the small set of COMMITTEES considered is a fraction of those that should be. That is, the larger set of COMMITTEES receiving reporting a combined total of $5 billion, versus the $740 million received by the 1997 COMMITTEES included in the CIR_3364 graph. Similarly, there are a total of $500 million in contributions coming from all COMMITTEES than these going to CANDIDATES, vs. the $158M from the smaller set. Obviously this set must be expanded.

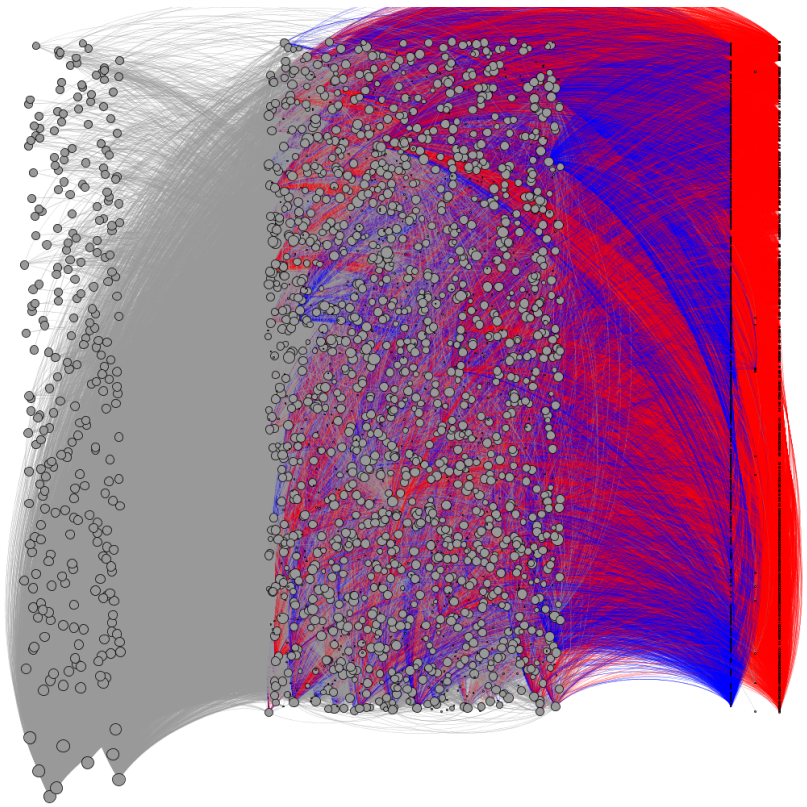

Figure 4: CIR map including all contributors

2 Mapping contribution flows

From this basic graph template, many different "maps" of contribution flows from specific nodes can be generated. Figure 4 shows a first graph placing CONTRIBUTORS in the far left column, CANDIDATES in the far right column and COMMITTEES between these.

The nine strata of non-BIGGIVERS appear as the nodes in the far lower left corner. Contributions from other sources appear along the left-hand side, ordered from biggest donors to smallest at the top. The smallest contribution to qualify as a distinguished bigGiver on this this list is around $150,000.

The CANDIDATE nodes are ordered with Presidential candidates at the bottom, through Senate candidates and then House candidates towards the top. Presidential candidates Obama and Romney have been pulled especially far down as they are obviously primary source of primary recipients of campaign country issues. Party affiliation of CANDIDATES associated with Democrats and Republicans have been colored blue and red, respectively, and edges to these nodes have been colored according to their target's color.

To simply further, removing all CONTRIBUTORS but except BIGGIVERS from the graph simplifies it considerably. The name associated with COMMITTEES can also be used to make some direct assignment of party affiliation with committees; e.g, the committee name "`NEVADA STATE DEMOCRATIC PARTY`" matches the regular expression "`.*DEMOCRAT.*`". This method can be used to color COMMITTEES nodes in the middle, as well as the edges with them as targets. This generates the map in Figure 5.

The final manipulation pulls those COMMITTEES affiliated with either Democrat or Republican by name (and colored in Figure 5) to the top and bottom. More clear here is how remaining nodes in the middle have strong party affiliation, without the party brand name in the title.

3 Contested races

A race, defined to be the set of CANDIDATES competing for the same office, is a much more narrow, locale-relevant use of this data. It extracts the subset of all contestants for a race, their contributing committees, and the contributors to these committees. This direct opposition between two candidates is arguably related to forms of "competitive coevolution" arising in other contexts, for example biological virus/host evolution and computational games.

In experiments below, we focus on a subset we will call Races84: the 84 Sen-

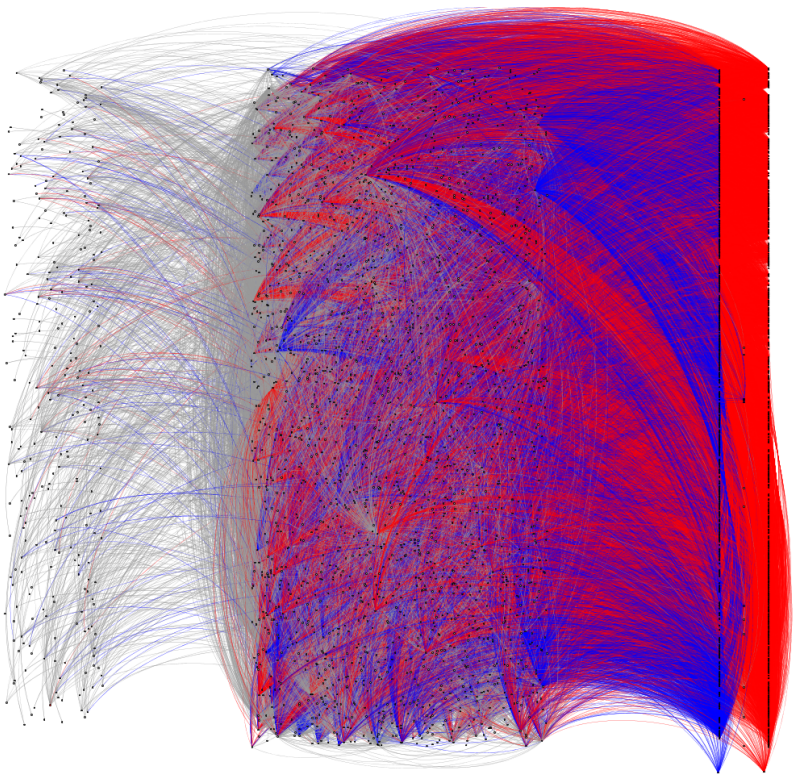

Figure 5: CIR big givers map

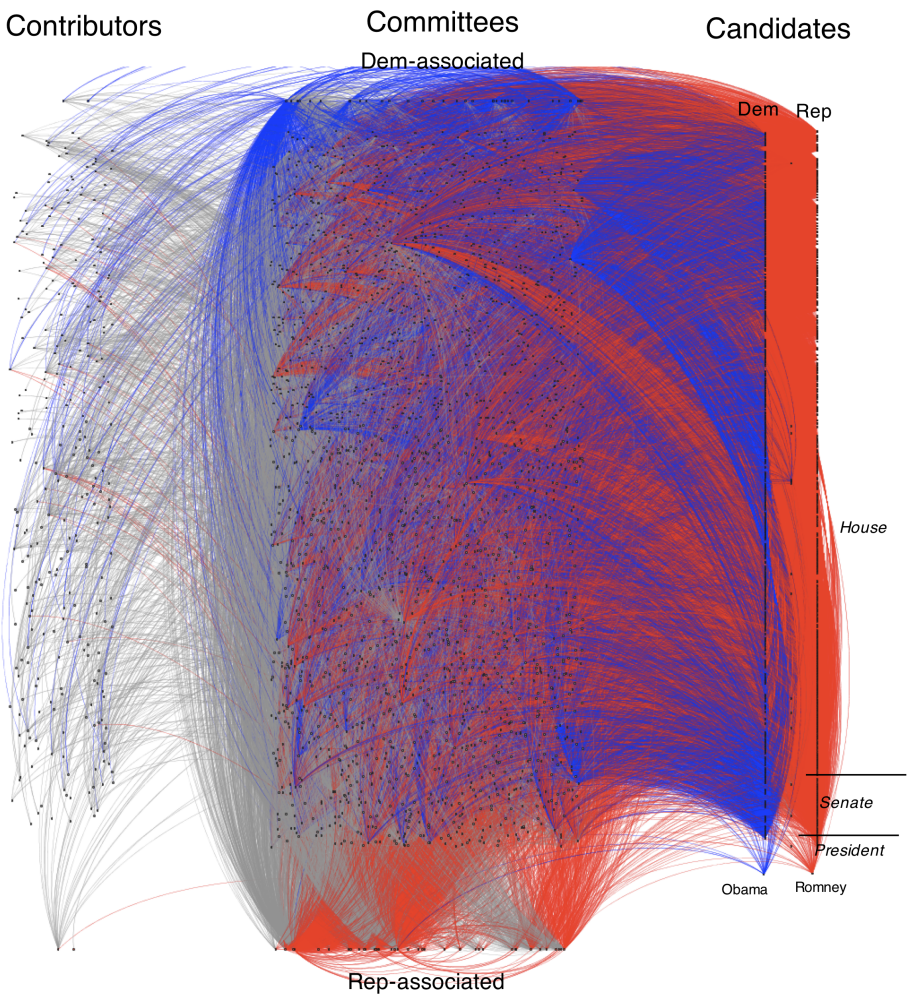

Figure 6: CIR map, annotated

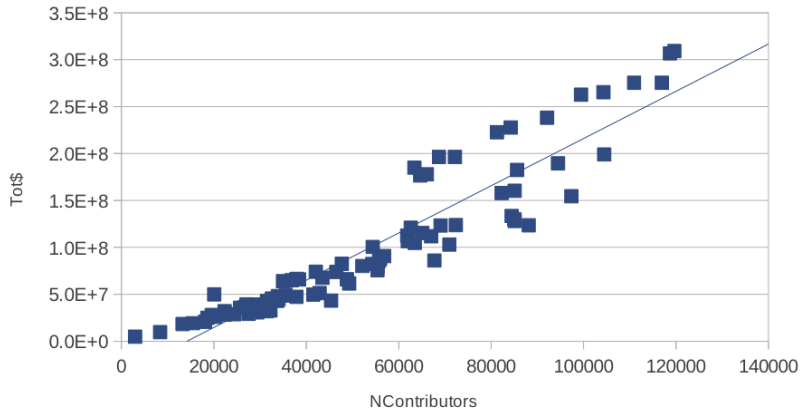

Figure 7: NContributors vs. Total$

ate and House races from the five states CA, MN, MA, MT, WI. Initial analysis distinguishes contributions coming to the candidates from the same state (as that within which the race is being run) from extra-state committees and contributors. Some patterns across this set are clear. For example, Figure 7 shows the high correlation of the number of contributors involved with the total numbers of dollars flowing into the race. While the general trend is not surprising, there appears to be an interesting change in the relationship at approx. 60000 contributors and $120M, from a very tight relationship below this threshold, to a more noisy one above.

Figure 8presents an example from a "typical" race, the California House District 1 race between DOUG LAMALFA(R) and JAMES REED (D). It is typical in that it poses one Democratic candidate against one Republican candidate; 52 of the 84 races are typical in this sense. The two candidates are represented with large nodes in the center, using the same Democrat:Blue::Republican:Red color convention as above. Below the candidates are two strata of node representing intra-California committees (nearer the candidates) and contributors (on the bottom row). Above the candidates are two much larger sets of nodes

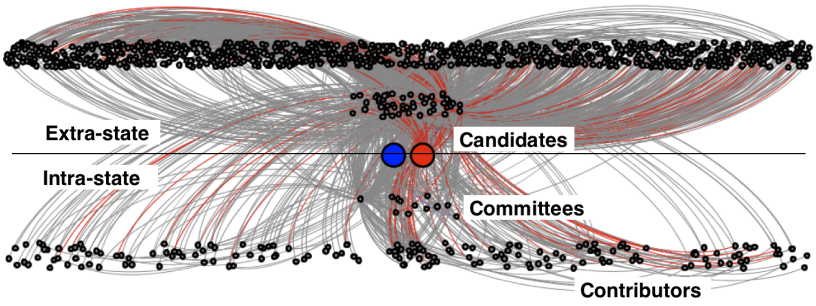

Figure 8: CA_House_01 race

representing committees (nearer the candidates) and contributors (on the top row) from outside California. In this particular race, there were 25643 contributors and 61 committees associated with either associated Lamalfa or Reed, but only the top 1000 contributors are included in Figure 8. This race ranks 67 out of 84 races in terms of the number of dollars contributed.

Certain giving patterns can be distinguished, for example from intra-state donors to intra-state committees vs. to extra-state committees, and from extra-state contributors to extra-state committees vs. to the candidates themselves. Other distinctions arise in contrast to other races, for example the CA House 2d race, shown in Figure 9. Here, the blue (Democratic) edges appear more dominant, just as the red (Republican) ones do in Figure 8. Note that the Democrat (Huffman) won in the 2d district, while the Republican (La Malfa) won in the 1st district.

As has been widely reported, California had it first elections under Proposition 14 of the 2010 election, the "top two primaries" initiative. This has created several races pitting two members of the same party against one another. One such race occurred in the California District 8 House race (Imus vs. Cook). This race ranked 83 out of 84 in terms of total dollars involved, but also has many more extra-state committees and contributors.

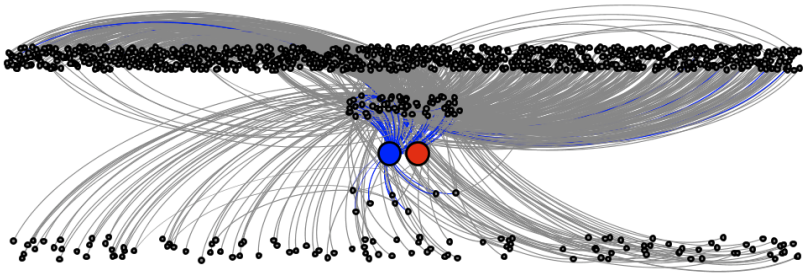

Figure 9: CA_House_02 race

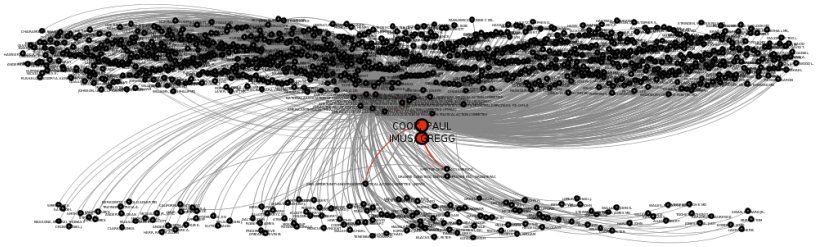

Figure 10: CA_House_08 race

4 Data cleansing efforts

I am humbled by my first foray into the FEC data. It seems a potentially profound window into the economic mechanisms underpinning modern politics in the U.S., but also apparently under-documented, and perhaps inconsistent in its data model and containing "dirty" data. I began to learn about this data only in September, 2012 and realize just how much more there is to know. I close with some of the ambiguous data features that seem most fundamental to further analysis, as well as some of the resources that I've come to most appreciate.

4.1 Anti-contributions

On its face, there is an odd pattern of Democrat CONTRIBUTORS/COMMITTEES giving to Republican COMMITTEE/CANDIDATES, and vice versa. A post on the Kaggle forum first identified the special character of some Type24 FEC contributions, as "anti" contributions. Similarly, *negative* donation amounts seem to also have to do with anti-contributions.

4.2 Normalization

Some aspects of data seem to require normalization that is typical in much other big data analysis:

1. Races: While there seems to be a convention in assigning unique candidate IDs of (e.g., H0KY05015="ROGERS, HAROLD DALLAS", a House candidate from Kentucky's 5th district), there are many examples where the candidates' IDs do not predict their office (e.g., H6CA07043="MILLER, GEORGE, from California's 11th district), and so automatically bringing all candidates involved in the same race together was not trivial. My solution was to get a sample of data from pages at OpenSecrets.org (retrieved 2 Nov 12) for the races comprising Races84.

2. Candidate names: Perhaps because the data I retrieved from OpenSecrets came from different sources than the FEC data, significant candidate name-normalization was required.

3. Entity_type: : I tumbled late to the potential significance of the entity_type attribute in the IND,ORG,CAN,PAC,CCM,PAC and COM data records. The logic underlying the distinctions made in Section 1.2 needs to be revisited to apply this data.

4.3 Resources

(I wish I had known about these from the beginning!)

- http://influenceexplorer.com/
  - I only ran across http://influenceexplorer.com/about/methodology/campaign_finance via a google search: "fec 24k pac"!?
- http://services.sunlightlabs.com/docs/Sunlight_Congress_API/
- http://www.followthemoney.org/index.phtml
- http://prototype.nytimes.com/gst/apitool/index.html
- http://www.programmableweb.com/apis/directory/1?apicat=Government