Federal Election Commission Campaign Data Analysis

This post is inspired by Marko Rodriguez' excellent post on a Graph-Based Movie Recommendation engine. I will use many of the same concepts that he describes in his post in order to load the data into Neo4J and then begin to analyze the data. This post will focus on the data loading. Follow-on posts will look at further analysis based on the relationships.

Background
The Federal Election Commission has made campaign contribution data publicly available for download here. The FEC has provided campaign finance maps on its home page. The Sunlight Foundation has created the Influence Explorer to provide similar analysis.

This post and follow-on posts will look at analyzing the Campaign Data using the graph database Neo4j, and the graph traversal language Gremlin. This post will go about showing the data preparation, the data modeling and then loading into Neo4J.

The FEC Data
The FEC data is available for download from the FEC website via FTP. It is composed of three main files which are the Campaign Committees, Campaign Candidates and the Individual Contributors. As of this post, there were approximately 10,875 committees, 3,600 candidates, and 455,000 unique contributions. Each of the data sets has a data description as well as frequency counts. The 2011-2012 data can be found here.

Gremlin and Neo4J
Gremlin 1.3 is available for download at this location. Neo4J 1.5M01 is available for download at this location. For this demonstration, we will be running the community edition of Neo4J in a Windows Virtual Machine.

Data Preparation
The FEC data is in formatted, fixed-length fields. This makes it a little bit harder to prepare for import into Neo4J with my limited skills and abilities. To work around that, I was able to load the data into Oracle using SQL Loader and then I wrote a simple PHP program to query the database and format the data into a delimited file. If interested in those files, feel free to contact me.

The FEC Data Graph
The FEC data is represented in the following graph. Each committee supports a candidate. Some candidates may be independent from a committee. Individuals contribute 1 or more times to a committee. For this demonstration, we've haven't separated out city/state/zip and created a common location.

A couple of notes on the data. Some of the committees did not have a treasurer so I added in a value of "No Treasurer". Some of the candidates were referenced to non-existent committees. In

this case, I've created entries for those committees in order to load the data and create the links. Additionally, the individual contribution file has overpunch characters to different amounts or negative amounts. Those values were adjusted in the database so the data could be loaded as an integer value.

Loading Data
The data will be inserted into the graph database Neo4j. The Gremlin/Groovy code below creates a new Neo4j graph, removes an unneeded default edge index, and sets the transaction buffer to 2500 mutations per commit.

Loading Committee Data
The committee data contains information about the different election committees. In our case, it has seven columns.

The code needed to parse this data is below:

Parsing Candidate Data
The candidate data contains information about the various candidates. In our case, it has nine columns. A sample of the data is below:

The code to parse the candidate file is:

Loading the Individual Contributors File
The individual contributors file contains all of the contributions made to different committees.

The sample data is:

Given that there are about a half a million contributors, parsing this data and loading will take a couple of minutes.

To commit any data left over in the transaction buffer, successfully stop the current transaction. Now the data is persisted to disk. If you plan on leaving the Gremlin console, be sure to g.shutdown() the graph first.

Validating the Data

Let's look at some distributions
What is the distribution of contributions among states?

What about the average contribution?

Are there any treasurers supporting multiple committees?

No chair and no treasurer indicate that the treasurer value was empty. However, there are several treasurers supporting multiple committees.

Next Steps
The next steps will be to look at some of the relationships between contributors and committees and see if there are treasurers serving on multiple committees.

Additionally, because each contribution is counted individually, there are several duplicate donors/campaign contributors. In order to address that, I will separate out the donors and their address as a separate table and link them to the contributions.

If you have questions about this post, feel free to email me.