2018 United States Primary Elections

An In Depth Analysis on Candidate Funding and Endorsements

Data 102: Data, Inference, and Decisions, Spring 2021

Jessie Houng, Ritvik Iyer, Sunny Shen, Rithika Neti

I. Introduction:

The United States has often been described as a political machine[1], a system by which small groups control and have command over election results. The disillusionment that many Americans feel today in regards to their elected officials comes from the increasingly predictable election results[2]; however, it is important to recognize that the predetermination of leading candidates or winning parties in a district stems directly from campaign financing and key party endorsements. Thus, we aim to more clearly understand the impacts that:

1. Preconceived partisan notions, quantified through a district's political lean, have on campaign strategy and popularity
2. Attributes such as endorsements and number of competitors has on raising and sustaining campaign funds

Since presidential election years can often skew campaign and election turnout, we primarily analyzed these impacts in the 2018 primary election cycle.

II. Data Overview:

This project mainly relies on 2 data sources: 2018 Primary Candidate Endorsement from FiveThirtyEight, and Campaign Financing Data in 2016 & 2018 from Federal Election Commission (FEC).

The 2018 Primary Candidate Endorsement Data

The 2018 Primary Candidate Endorsement data is the census since it includes all candidates who received votes in the Primary 2018 election. The data includes people who ran for the U.S. Senate, U.S. House and governor in which no incumbent ran, so races that had an incumbent are systematically excluded from the dataset. The dataset contains information about candidates' personal information, endorsements, and the outcome of elections. For personal information fields such as gender, race, and whether they identified themselves as LGBTQ+, the values came from candidates' websites or their public speeches, and the fields were filled with NaN if the candidate never explicitly identified themselves with a specific category. The values about endorsements are gathered from each person's or organization's endorsement/anti-endorsement lists, whereas the data for election outcomes are mostly supplied by Ballotpedia. All the information is either publicly available online or made public by the candidates themselves, so they should be well aware of the collection, spread, and use of these data.

In the dataframe, each row corresponds to one candidate, which convinces us to use the candidate's information (name, state, party) as the unique identifier when we merge the data with other datasets. This also motivates us to examine whether there are patterns that exist across all candidates and what we may be able to predict for a random candidate given their personal and campaign information. We hope to generalize our findings to the primary election in 2018 in general, but one concern is that the data does not contain races featuring incumbents

[1] DiGaetano, Alan. "The Origins of Urban Political Machines in the United States: A Comparative Perspective." *Urban Affairs Quarterly*, vol. 26, no. 3, Mar. 1991, pp. 324–353, doi:10.1177/004208169102600302.

[2] Aaron Martin, Zoe B. Pidgeon, Robert Calimente, Alexandra D' Antonio & Albi Taipi | Richard Meissner (Reviewing editor) (2019) A new campaign strategy informed by pragmatism: Running on a platform of expanding voting accessibility, Cogent Social Sciences, 5:1, doi: 10.1080/23311886.2019.1631526

and those races might look significantly different than the rest of races. Another concern about the dataset is that values of some columns are gathered from interviews or indirect references by the FiveThirtyEight team, and is therefore subject to human errors. However, we believe that there should be a very small amount of errors (if any) as FiveThirtyEight is a well-organized website, so it should not have a significant impact on our analysis. We wish there were a column indicating how progressive or conservative a candidate is so we could see how different scores on the political spectrum influence election results, endorsement, fundraisings, etc. We only have a "Partisan lean" column, which is calculated by the difference between how a state or district voted in the past two presidential elections and how the country voted, to study whether the Partisan lean of the district that a candidate comes from has any impact on their campaign.

Campaign Financing Data in 2016 & 2018

We downloaded 3 datasets from the FEC data:

1) The All Candidate summary file that contains one row for each candidate with summary financial information
2) The Candidate-Committee linkages file that contains information linking the candidate's information to information about their committee.
3) The Individual Contribution data from 2015-2018 from the FEC website.

All three datasets are the census as candidates are legally required to report to the FEC about their financing. However, it only includes complete data for candidates running for seats in the House & Senate and not candidates running for governors, as governors are not legally obligated to report. All the candidates should be aware of the collection of the data, but the individual Contribution data also includes information of the individual contributors such as their names, amount donated, zip code, employer, and occupation. It is unclear to what extent that the individuals are aware that those personal information will be collected and made public on the FEC data.

In individual Contribution data, each row in the dataset represents each transaction. Therefore, we will need to find unique identifiers for individual contributors in order to see how each person contributed. We wish there were a unique identifier for each person in the dataset, but there is only a unique identifier for transactions. As a result, we need to find other ways of constructing identifiers for individuals, which will be explained in the Data Cleaning section. It is unclear whether all the contributions data are collected digitally, or some of them may have been collected through paperwork and later entered into the database by people, which again is subject to human errors. In this case, we will assume that human errors (if any) are negligible and will not influence our analysis.

III. Data Cleaning:

The 2018 Primary Candidate Endorsement Data

As the FEC data only includes candidates running for Senate and House not candidates running for governors, we need to drop all the rows in the Candidate Endorsement dataset that contain information about candidates running for governors. The Candidate Endorsement dataset contains two csv files: one for Democratic candidates and one for Republican candidates. We added a column indicating which party they are from and then concatenating the two dataframes.

There are a lot of NaN values in Endorsement columns, which is expected because "Yes" refers to the person/organization explicitly endorsing the candidate, "No" refers to the person/organization explicitly

anti-endorsing, and “NaN” if no information about the attitude of the person/organization is available. Therefore, we replace “Yes” with 1, “No” with -1, and “NaN” with 0 to indicate the positive/negative/neutral endorsement status.

Only the Democratic csv contains information about Partisan Lean, which is calculated by the difference between how a state or district voted in the past two presidential elections and how the country voted. Therefore, we extracted the Partisan Lean value from each district and input it into the Repblican rows so that most rows in the merged dataset have Partisan Lean information. There are, however, some states in which Republican candidates still have missing partisan lean values because no Democrat candidates ran in those districts/states and therefore no values are available to be input into those Republican rows.

We believe that the partisan lean of a district -- how voters behave in the district -- may have a significant impact on how much support a candidate gets. We dropped the rows that do not contain partisan information, which is about 19% of the rows in the Republican dataset and a significant portion. In order to proceed with our analysis, one important assumption we need to make is that for rows that have partisan lean information and the rows that don’t, there is no significant difference of races and support for candidates between the 2 groups. We will evaluate this assumption in our first research question.

Campaign Financing Data in 2016 & 2018

It is easy to merge the 3 FEC datasets together as a unique candidate ID is provided as the primary key. However, there is no unique identifier for candidates in the Endorsement Data. We use the candidate name, state, and party affiliation in both datasets to merge the Endorsement Data and FEC Data. We observed low matching cases at first because the formatting of the names are different. For example, “Tim Westley” in the Endorsement Data is “Timmy Lee Westley” in the FEC data, and “Ed Meier” in the Endorsement Data is “Edward Meier” in the FEC data. After researching string matching methods, we decided to use the fuzzy_match package to match names in both datasets, and we ended up successfully matching 70% of the candidates that ran for House of Senate in the 2018 Primary Election.

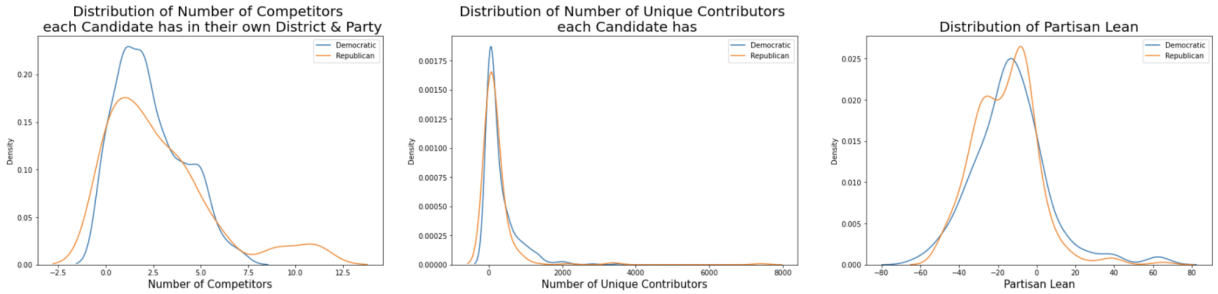
Both the Candidate-Committee linkages files and the Individual Contributions files contain information about all kinds of elections and contributions that occurred from 2015-2018. Therefore, we only select rows corresponding to Primary Elections in 2018. In the Individual Contribution data, we use Name + Committee + Zipcode to identify unique contributors to a committee. We merge the two dataset based on Committee IDs, and then group by Committee IDs and Candidates to calculate the number of individual contributions each candidate receives (in all their committees) in the Primary Election of 2018. We then repeat this process for the Primary Election in 2016 so that we can use it as a prior for later analysis.

After all the data cleaning is done, we merge Endorsement Data and Campaign Financing Data together. The final dataframe has one row for each candidate with their personal information, state and district that they are running in, party affiliation, election results, endorsements, and number of individual contributions.

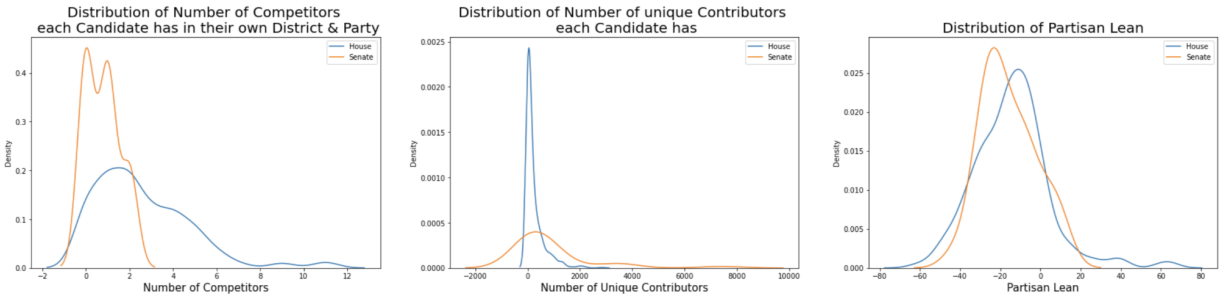
IV. Data Exploration:

We are curious about whether there are any significant differences between the Republican and Democratic Party in terms of how many unique individual contributions they get, and how many competitors from the same party and the same district they have, and the partisan lean of districts that the candidates are from. In the

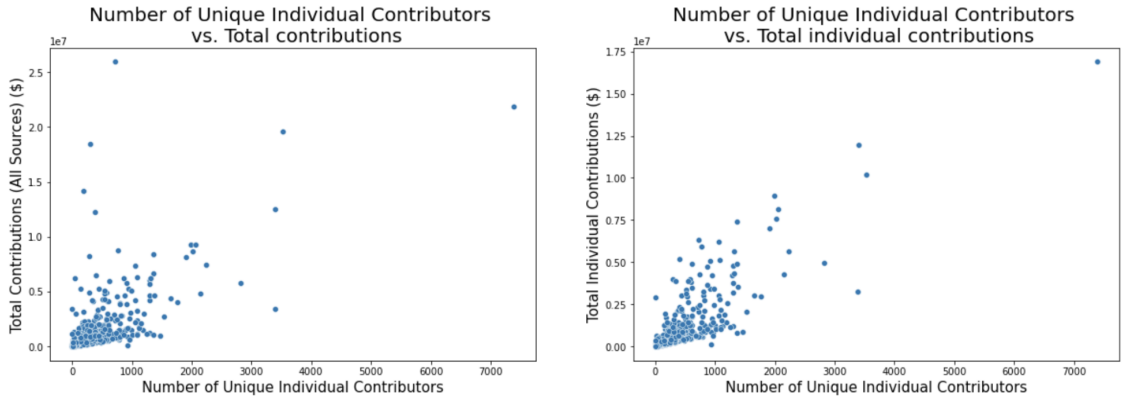
overlaid distplots below, it seems like that the distribution of the number of individual contributors each candidate had and the distribution of partisan lean look roughly the same for both parties. The distribution of the number of competitors in both parties follows a similar pattern, where most people have 0-8 competitors, although Republican candidates are more likely to get a larger number (>10) of competitors from the same party and the same district.



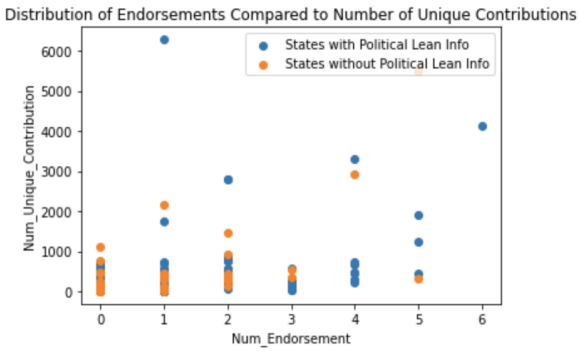
While the Party affiliation does not seem to make a big difference in those distributions, we see that Office Type -- House Representative vs Senator -- makes a big difference in the distribution of number of competitors and number of unique contributors. Candidates for house representatives tend to have more competitors, and candidates for Senators tend to receive a larger number of individual contributions

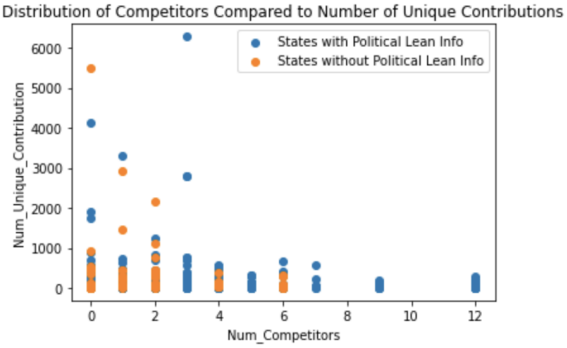


We also investigate the relationship between the number of individual contributors and the total contribution received by a candidate from all sources, and between the number of individual contributors and the total individual-only contribution received by a candidate. While both graphs show positive linear relationships, we see a few outliers in the graph on the left-hand side that some candidates did not have a lot of individual contributors but raised a lot of funds, which suggests that other money sources (i.e. contributions from political committees) is an important factor too.



In addition, we wanted to compare the distributions of various factors in order to better understand the differences between states that reported partisan lean information and those that did not. We found that while the distributions looked similar, we wanted to test and ensure that we would be able to treat both sets of states equally in our analysis. This guided our first research question of understanding the key differences between states that reported partisan lean information and states that did not.





V. Research Questions:

Our data cleaning and exploratory data analysis guided the following research questions:

1. What are the differences between candidates who have missing partisan leans and those who do not have missing partisan leans? Are these differences significant?

2. How does the political lean of the district, number of positive endorsements, and number of competitors influence the number of unique donations a candidate receives in primary elections?

V.i Multiple Hypothesis Testing:

The primary purpose of our multiple hypotheses testing is to determine whether we are inherently biasing our results by dropping candidates with missing partisan lean data. Along with this, we hope to better understand whether states with missing partisan lean data contain significantly different electoral campaign characteristics, potentially influencing voters beliefs. To do this, we compare distributions of the subsets of our data containing candidates with and without known partisan lean information, across several features.

Methods:

One key observation to note is that partisan lean is defined as a measure of how strongly a congressional district leans towards the Democratic or Republican Party, meaning that if we do not have the partisan lean for one row (candidate), we will not have the partisan lean for every candidate in their district. This made us hesitant about dropping rows without partisan lean data, as it would mean excluding entire districts in our analysis. Thus, we decided to group our dataset by district and run hypotheses tests to determine how excluding rows with incomplete information would affect our overall distribution. Specifically, we focus on the distributions of:

1. The number of candidates in each district
2. The number of positive endorsements candidates of each district receive overall

Additionally, we split these two distributions by the type of office position the candidates are running under (‘Representative’ or ‘Senator’), to further analyze whether the distribution relationships hold. If we had analyzed candidates as independent rows, our distributions for our features, specifically endorsements, would have very little variation, as people are seldom endorsed in general, especially in these primary elections. For each given endorser in the data, realistically they probably would not endorse multiple candidates especially if they run against each other, or are in a district they have no tie to (for example Andrew Yang is from New York, so he would give endorsements to NY candidates so it would be more impactful in those districts). Ultimately, our intuition is that if the distribution of valid (known partisan lean) candidates is statistically similar to the full data in the features we study, then we can conclude we are not excluding significantly impactful data.

Beyond looking at the district-specific distribution, we wanted to ensure that between the states that reported partisan lean and those that did not, the key associations using the number of contributions did not differ significantly, further evidencing that we could not only drop the rows but also demonstrate that preexisting partisan information is not a cause of differences in campaign financing or establishment endorsements. For these tests, all candidates were grouped together as it allowed for the distributions to be compared more broadly and ensured a larger dataset. Specifically, we compared the distributions across the two sets of states (those that reported partisan lean and those that did not) through the associations of number of contributions, number of contributions and number of endorsements, and number of contributions and number of competitors. The three association tests that were performed comparing the two sets of distributions included:

1. Solely comparing number of contributions - This compares the distribution of the number of contributions across both sets of states. This was studied in order to whether the disclosure of partisan lean information impacted the number of donors a candidate got, thus demonstrating whether preexisting partisan information alters donors.
2. Comparing the association between the number of contributions and the number of endorsements - This compares the distribution of the association between the number of endorsements and the number of contributions for both sets of states. This was studied in order to guarantee that the correlation between the two variables was consistent regardless of partisan lean information.

3. Comparing the association between the number of contributions and the number of competitors - This compares the distribution of the association between the number of competitors and the number of contributions for both sets of states. This was studied in order to guarantee that the correlation between the two variables was consistent regardless of partisan lean information. For this association, the values were grouped by state in order to have less varied and more consistent data.

In terms of carrying out our hypothesis tests, we choose to do a two-sided hypothesis test where we compared two distributions. Therefore, broadly, our null hypothesis stated that the two distributions that we were comparing were roughly the same; whereas, our alternative hypothesis stated that the two distributions were significantly different.

For our test statistic, we decided on utilizing the KS-statistic from the two sample Kolmogorov-Smirnov (KS) test-- a non-parametric test that compares the differences in the empirical cumulative distribution functions (ecdfs) of two samples. Because we specifically wanted to compare two mutually exclusive subsets of the data (districts do not overlap between valid and invalid subsets), the two sample permutation KS test was a fitting choice. We used a permutation test here because the original KS-test requires both samples to come from continuous distributions. The KS statistic is sensitive to both location and shape differences in the given distributions, which makes it appropriate in this context.

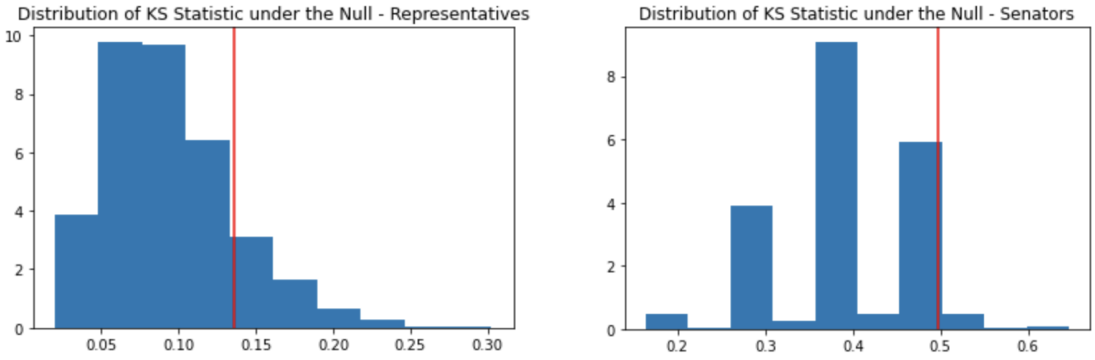
We know from our data that the valid and invalid distributions are from the same underlying distribution (the null hypothesis is true), so we want to mitigate the false discovery rate (the case that we reject the null hypothesis when it is in fact true). Essentially, we looked to decrease the probability of falsely rejecting the null hypothesis. Thus, we looked to apply two forms of correction procedures to do so—Bonferroni and Benjamini-Hochberg.

Our first attempt at correcting errors in our decisions was with the Bonferroni procedure, where we essentially modify our decision rule to comparing p values to a significance level of α/m, where α is an overall desired value, and m is the number of hypotheses being tested (in our analysis $m = 7$). Contrary to our goal, the new decision cutoff was higher than .05 (our default cutoff rule for 95% confidence), meaning that the p values were more likely to be rejected under the null hypothesis. Intuitively, we recognize that the Bonferroni procedure is more suited for controlling the familywise error rate (FWER), which is the probability of making at least one false discovery in our hypothesis tests. Ultimately, to correct our decision errors, we applied the Benjamini-Hochberg procedure in order to control the FDR. Contrary to the Bonferroni procedure, Benjamini-Hochberg essentially makes it more likely to fail to reject the null hypothesis (decision rules in favor of the alternative), and overall decreases the probability of falsely rejecting the null hypothesis in our tests.

Results:

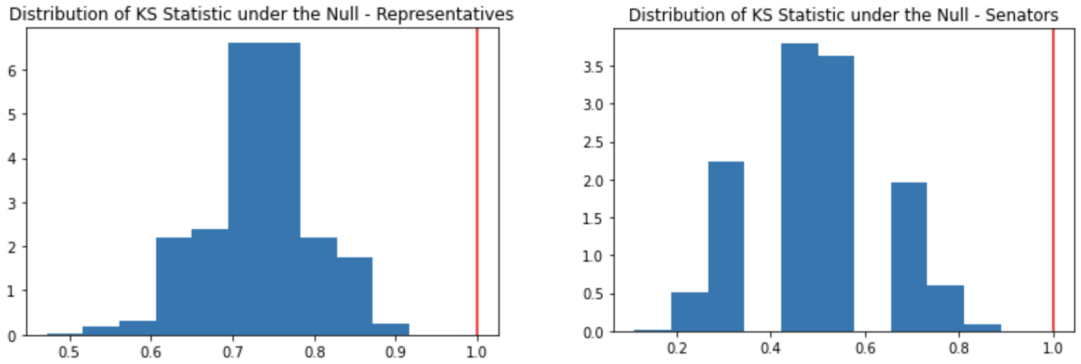
Testing:	Null Hypothesis:	Alternative Hypothesis:	P-value	Decision
Candidate Counts for Representative Office	Candidate counts for our Representative office valid and invalid subsets are from the same	Candidate counts for our Representative office valid and invalid subsets are not from the same	0.2712	Fail to reject the null

	underlying distribution.	underlying distribution.		
Candidate Counts for Senator Office	Candidate counts for our Senator office valid and invalid subsets are from the same underlying distribution.	Candidate counts for our Senator office valid and invalid subsets are not from the same underlying distribution.	0.596	Fail to reject the null
Number of Positive Endorsements for Representative Office	The number of positive endorsements for our Representative office valid and invalid subsets are from the same underlying distribution.	The number of positive endorsements for our Representative office valid and invalid subsets are not from the same underlying distribution.	0.0	Reject the null
Number of Positive Endorsements for Senator Office	The number of positive endorsements for our Senator office valid and invalid subsets are from the same underlying distribution.	The number of positive endorsements for our Senator office valid and invalid subsets are not from the same underlying distribution.	0.0	Reject the null
Number of Contributions	The distribution of the number of unique contributions per district is roughly the same in states that have reported partisan lean info and states that do not.	The distribution is significantly different.	0.1942	Fail to reject the null
Association Between # of Endorsement and # of Contributions	The distribution of the number of contributions based on endorsement is the same between the subsets of data with and without missing partisan leans	The distributions are significantly different	0.713	Fail to reject the null
Association Between # of Competitors and # of Contributions	The distribution of the number of contributions based on competitors is the same between the subsets of data with and without missing partisan leans	The distributions are significantly different	0.0	Reject the null



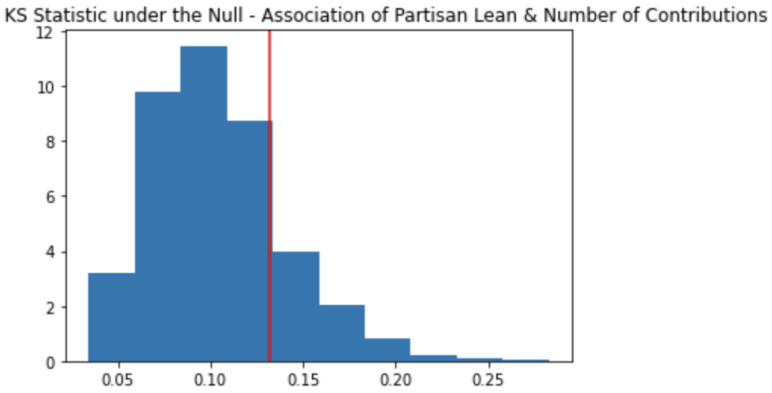
left: fig 1.1, Distribution of KS Statistic (candidate count) under Null for Representative runs; right: fig 1.2, Distribution of KS Statistic (candidate count) under Null for Senator runs

In fig 1.1 and fig 1.2, we demonstrate the distribution of the KS statistics generated on candidate count by district distributions for candidates running for Representative and Senator offices respectively. The respective p values (red line) are strongly significantly above the p value cutoff, even prior to correction with Benjamini-Hochberg. Therefore, from our test results we can conclude that removing the rows with missing partisan lead data will not significantly affect the overall distribution of candidate counts.



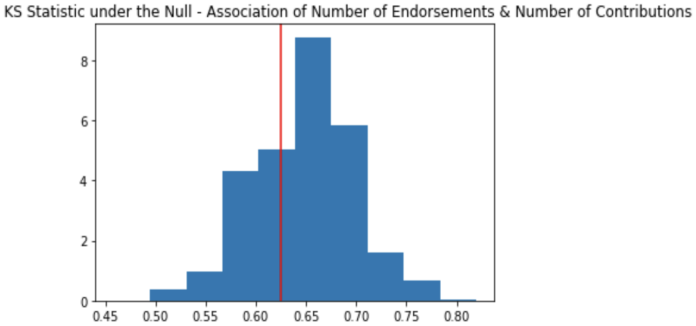
left: fig 1.3, Distribution of KS Statistic (positive endorsement count) under Null for Representative runs; right: fig 1.4, Distribution of KS Statistic (positive endorsement count) under Null for Senator runs

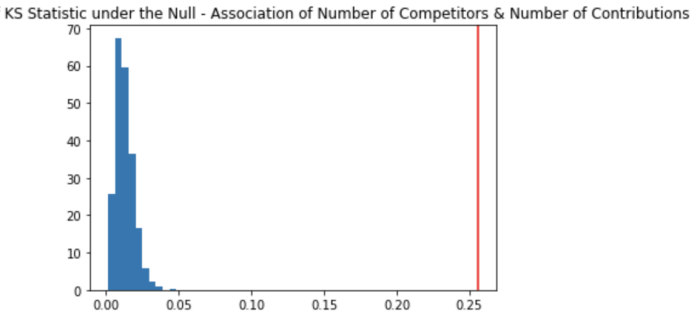
In fig 1.3 and fig 1.4, we demonstrate the distribution of the KS statistics generated on the number of positive endorsement per district distributions for candidates running for Representative and Senator offices respectively. The respective p values (red line) are strongly significantly below the p value cutoff, even after correction with Benjamini-Hochberg (the p values are essentially 0 in both cases). Therefore, from our test results we can conclude that removing the rows with missing partisan lead data will significantly affect the overall distribution of positive endorsement counts. However, the number of positive endorsements received is very low across all candidates and subsequently districts. We note that by excluding the rows with missing partisan leans we are introducing a source of bias and error in our remaining dataset for endorsement counts, and proceed the analysis with caution.



center: fig 1.5, Distribution of KS Statistic under Null for Number of Contributions

Figure 1.5 compared the distribution of the number of contributions for both sets of states. The distribution above demonstrates the results under the null using the KS statistic test. Based on the distribution plot as well as the significant p-value of 0.1942 prior to correction, it is clear that the distribution of the number of contributions is consistent across states that report their partisan lean and those that do not. Therefore, from this test result, we can conclude that removing the rows with missing partisan lead data will not significantly affect the overall distribution of candidate counts.





left: fig 1.6, Distribution of KS Statistic under Null for Endorsements & Contributions;
right: fig 1.7, Distribution of KS Statistic under Null for Competitors & Contributions

The tests carried out in figures 1.6 and 1.7 compared the association between two variables for both sets of states. The distributions above demonstrate the results under the null using the KS statistic test. For figure 1.6, the distribution compared the association between the number of endorsements and the number of contributions. Based on the distribution plot as well as the very significant p-value of 0.713 prior to correction, it is clear that the distribution of the number of contributions is consistent across states that report their partisan lean and those that do not. Therefore, from this test result, we can conclude that removing the rows with missing partisan lead data will not significantly affect the overall distribution of candidate counts. For figure 1.7, the distribution compared the association between the number of competitors and the number of contributions grouped by state for less varied data. Based on the distribution plot as well as the p-value of 0, it is clear that even after correction, the distributions across the two sets of states are different. However, it is important to note that the number of competitors proved to be an inconsistent metric when measured collectively per state or district. The correlation between competitors and contributions generally was weak for both sets of data.

Based on the results of the multiple hypothesis tests, we failed to reject the null hypothesis on four out of the seven tests, meaning that the majority of tests demonstrated that the key factor distributions between states that reported partisan lean and those that did not were the same. This justifies our decision to drop the candidates with missing partisan lean information under the assumption that they are not significantly different from those candidates with this provided information.

Discussion:

After analyzing our results from the hypotheses tests, we attempted to control our FDR on the basis that we know the null hypotheses are true. We attempted both Bonferroni and Benjamini-Hochberg procedures to increase the likelihood of accepting our null hypothesis, but our results did not affect the number of p values we fail to reject at any chosen significance level input. Although we were hopeful and diligent about our correction methods, the p values we rejected were essentially 0, meaning that any decision rule that would end up failing to reject them would be impractical and frankly misleading.

Synthesizing our methods, results, and correction attempts, we have concluded that dropping the candidates with missing partisan lean information will not significantly impact further analysis. We recognize that because we still reject three of our null hypotheses after correction, we must be cautious about the potential bias we introduce when dropping rows from our dataset. However, overall we have enough evidence to assume that the variables we consider indicative of the number of unique donations are not affected by these potential biases. Moreover, this question aimed to understand the impacts of preconceived partisan notions; for example, we wanted to understand if donors would be influenced to donate more in districts that did not tell them specifically what the party breakdown was. Some limitations and confounding variables in this data include:

1. Bias based on location of candidate. Despite missing partisan lean information, there are some inherent party leans in some of the states that were missing candidate information. For instance, Massachusetts was a state that had missing partisan lean information; however, the state has three times as many Democrats compared to Republicans[3].
2. Demographic factors. Beyond location, a candidate's demographic background influences their ability to raise money from specific donors. For example, during Hillary Clinton's presidential campaign, they received significantly more donations from women than previous elections[4].
3. External factors. Along with partisan lean, there are lots of different factors that can influence both endorsements, competitors, and contributions. With competitors, oftentimes, there will be fewer competitors depending on the experience of the incumbent or leading candidate. However, if there is no clear frontrunner, that can often lead to more competitors and less endorsements as key figures choose to "wait and see" before providing their endorsement. In this scenario, type of race would also have a large impact.

Looking ahead, this type of analysis could be used to inspect the impacts of polarization on a district or a particular candidate. Thus, it would be interesting to compare these metrics across various elections, including presidential election years. More detailed analysis of where the contributions are coming from such as donor location, previous donations, and demographic could be helpful in generating a more specific understanding of partisan lean and its impacts on other campaign-related factors.

---

[3] Elections: Massachusetts Registered Voter Enrollment

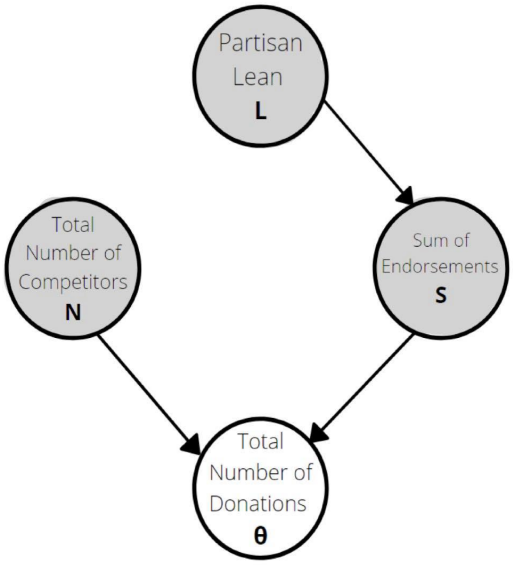
[4] Hillary Clinton's Female Donors. The Atlantic. 2016 January 1.

V.ii Bayesian Inference:

After showing in Q1 that key factor distributions for candidates with missing partisan lean information are not significantly different from those for candidates with partisan lean information, we would like to use our dataset to make inferences regarding grassroots fundraising in Primary 2018 campaigns. Specifically, we are interested in modeling the distribution of unique individual donations, which broadly provides insights into a candidate's political ideology, voter demographic, and campaign platform. Bayesian inference is especially relevant in the context of campaign fundraising because there is a notion of a "belief" or historical prior of previous election cycles, which can be updated using new candidate or campaign characteristics.

Methods:

Using this framework, we propose the following graphical model:

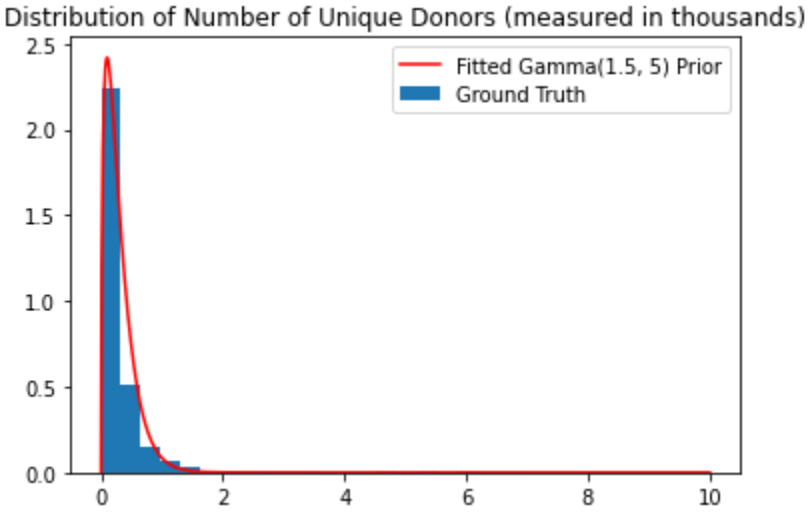


center: fig 2.1, Bayesian Hierarchical Model

In the figure above, we specify a graphical representation of a Bayesian Hierarchical Model, which aims to find the posterior distribution of the total number of unique donations,θ, given information on the partisan lean of a district L, the total number of competitors in a district N, and the sum of positive endorsements for a candidate S . We chose these variables in particular because they provide a general picture of a candidate's campaign success without introducing an extreme level of granularity in the model (e.g. a state/geographical variable). In a real-world scenario, this model would be fitted without knowing the ground-truth θ, so we regarded the total number of donations as the hidden variable, while the other variables are observed directly (indicated by shading).

Based on the hierarchical structure present in the model, the posterior probability distribution for θ can be expressed as: $p(\theta|L,S,N) \propto p(N|\theta)\,p(S|L,\theta)\,p(L|\theta)\,p(\theta)$, where $p(N|\theta)\,p(S|L,\theta)\,p(L|\theta)$ represents the likelihood function and $p(\theta)$ represents the prior distribution. First, we estimated the prior distribution using Primary 2016 data from the FEC. Using the individual contributor data of 2014 and 2016 as well as the corresponding committee-campaigner linkages dataset, we were able to obtain the unique donor counts for each candidate running in the 2016 Primary election cycle. After visualizing the data, we realized that there were large numbers of extreme values at the right tail, which made it difficult to fit a model accurately and estimate

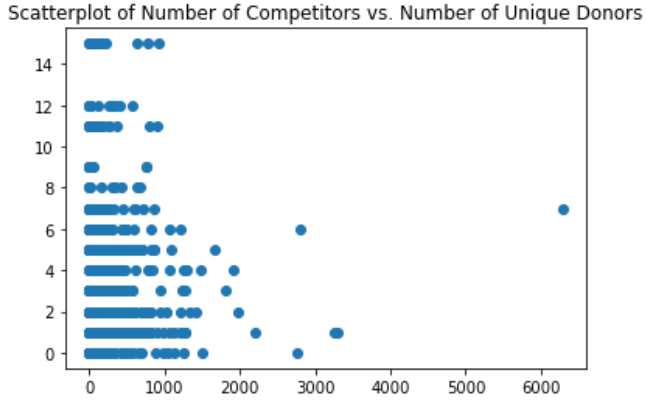
parameters. Therefore, we decided to prune extreme outliers (e.g. popular Senate candidates who had 100,000+ unique donors) and measure θin thousands of donors, instead of its original units. Implementing these changes lead to the following distribution:

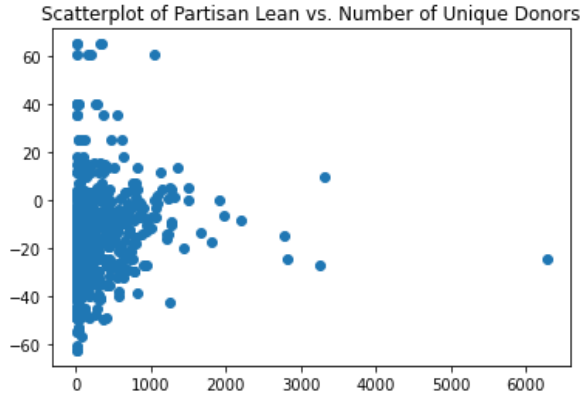


center: fig 2.2, Gamma Prior on 2016 Primary donor distribution

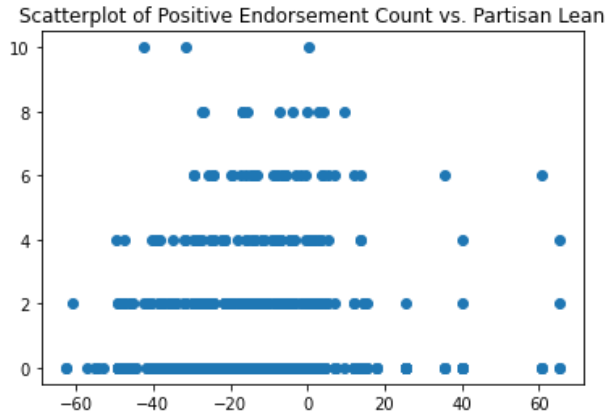
A main characteristic of the donor distribution is the strong right skew, which we were able to capture using a $Gamma(\alpha = 1.5, \beta = 5)$ distribution. Despite being a continuous distribution, the Gamma distribution is an appropriate choice for a prior since the units of θhave been re-scaled to measuring thousands of donors, so its distribution can assume decimal values. Overall, this prior distribution captures the belief that most candidates running in Primary elections will garner donations from somewhere between 0 and 1000 unique donors.

Next, we examined scatterplots of variables to infer the conditional distributions forming the likelihood.





*left: fig 2.3, Relationship between Number of Competitors (N) vs. Number of Unique Donors (*θ**);**
right: fig 2.4, Relationship between Partisan Lean (L) and Number of Unique Donors (θ**)**



center: fig 2.5, Relationship between Positive Endorsement Count (S) and Partisan Lean (*L*)

From the scatter plot in figure 2.3, we noticed that for small values of θ, the conditional distribution $p(N|\theta)$ appears to be fairly uniform on $[0, 15]$. However, we noticed that for larger values of θ, the points remained fairly evenly spread, but on a smaller interval. Therefore, we inferred that:

$$p(N|\theta) \sim DiscreteUniform[0,\ 15 - \theta]$$

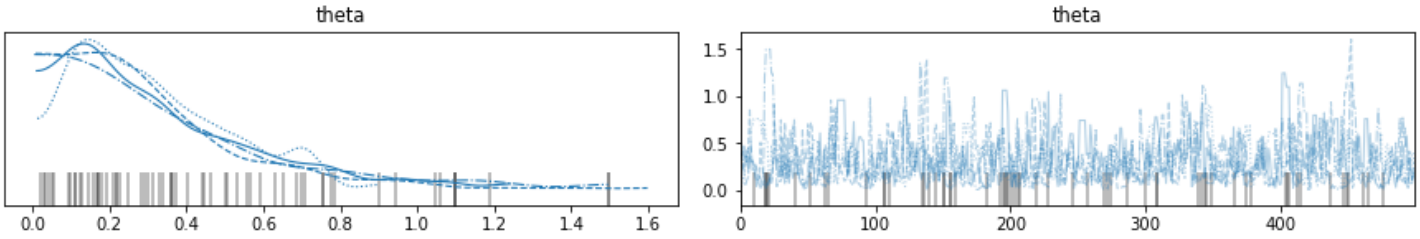
We took a similar approach to determining the conditional distributions for the two other key conditional distributions and inferred that:

$$p(L|\theta) \sim N(0,\ \frac{1}{\theta})$$

$$p(S|L,\theta) \sim DiscreteUniform[0,\ 10 - \frac{|L|}{10}]$$

We were careful to use non-negative, discrete distributions to model *N*and *S*because they represent counts. Finally, after identifying all the components to assemble the posterior distribution, we opted to perform inference via Markov Chain Monte Carlo (MCMC), as conjugacy does not hold for $p(\theta|L,S,N)$. Using the pyMC3 library, we used a tuning period of 4,000 samples on 4 chains, each drawing 1000 samples, for a total of 4000 samples to form our approximate posterior distribution.

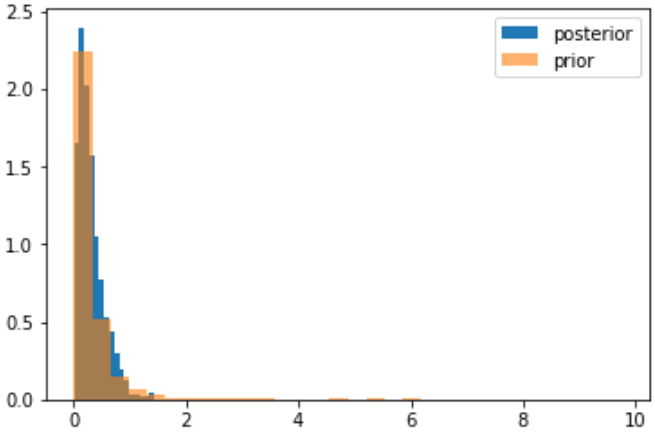
Results:

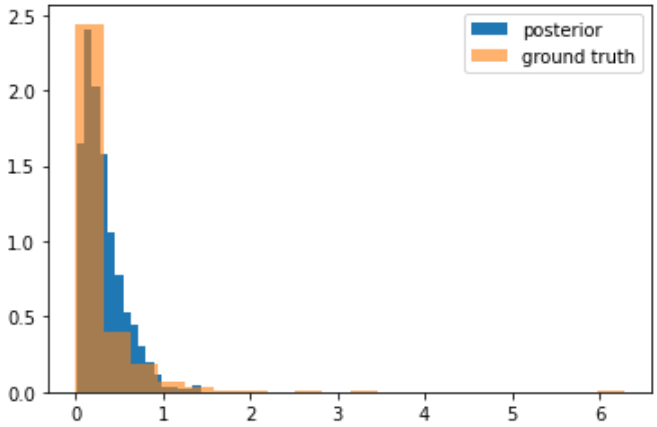


left: fig 2.6, Kernel Density Estimate of the Posterior distribution;
right: fig 2.7, Traceplot of the Posterior distribution

Figure 2.6 visualizes the shape of the posterior distribution, with the embedded rug plot showing areas of high density around 0.2 to 0.4. Therefore, the posterior distribution reflects that it is quite probable that a candidate in the 2018 Primary cycle will command between 200 and 400 unique donors. The figure on the right represents the sampled values over time, which were fairly variable, between 0.2 and 1.5 thousands of unique donors per candidate.

We were also interested in estimating the impact of the observed data in relation to the prior. To study this, we constructed the following plots:





left: fig 2.8, Overlaid Histograms of the Prior and Posterior distributions
right: fig 2.9, Overlaid Histograms of the Posterior and 2018 Primary Ground Truth

From Figure 2.8, we can observe that the simulated posterior distribution is more dispersed than the prior, so the observed data seemed to provide evidence that candidates in the 2018 cycle were more likely to draw larger numbers of donors. However, this idea is contradicted by the plot in Figure 2.9, which shows that the simulated posterior is indeed overdispersed compared to the true 2018 Primary distribution. The ground truth data has a larger concentration of candidates raising between 0 and 1 thousand unique donors, with a right tail which decays very quickly.

Finally, to capture the uncertainty for our estimation, we generated a 95% credible interval using the simulated posterior distribution. We found that, if our model is good, the random variable θ lies within [29.23001681, 944.75419883] (expressed in original units) with a 95% chance. Therefore, we can see our model finds that it is very likely for a typical candidate in the 2018 Primary election to receive between 29 and 945 unique donors up to the Primary election date.

Discussion:

There are several areas to improve in our modeling process. First, due to limitations in the older FEC data, we were unable to place priors separately based on campaign type (e.g. Senate vs. House). Including this variable as a hyperparameter would help to reduce the outlier effect we mentioned earlier and would produce better fitting priors. This is important because Senate campaigns are run on a state-level, which means that Senate candidates typically have much larger numbers of individual donations due to larger campaign-scale. At the same time, there are also less Senate candidates running in general, so it wouldn't make sense to pool them in with the rest of the candidates running for a House seat.

Additionally, the method by which we chose conditional distributions for likelihood was guided by intuition rather than by mathematical reasoning. An alternative approach would be to fit a probabilistic classifier to the data, which would learn the optimal conditional probability distributions (e.g. Naive Bayes). These improved likelihood functions would better represent the true data-generating process and allow for an improved posterior estimation. Finally, the model can be refined to include more variables like party and election cycle, so the results could potentially generalize to new election cycles as opposed to just the 2018 primary elections.

With respect to the inference procedure, there were around 150 instances of divergences which lead to low-quality posterior samples. This likely occurred because of the high acceptance target that we set of 0.95 and the difficulty of sampling from the non-standard conditional probability distributions we used to construct the likelihood. For example, $p(S|L, \theta)$ involves an absolute value, which is not completely differentiable, which could potentially pose problems for the underlying Hamiltonian Monte Carlo algorithm, which uses gradients to converge on the target distribution faster.

VI. Conclusion:

This paper aims to understand how factors like preconceived partisan notions, endorsements influence the number of individual contributors each candidate gets in the 2018 Primary Election. We merge the Endorsement data about the Primary 2018 election and Financing data from the FEC website so that we can connect the Endorsement and Finance aspects of elections. There are some missing values in the Partisan Lean, and therefore we conduct multiple hypothesis testing to see whether dropping those rows would impact our analysis. We conclude that the distribution of variables that we are interested in examining do not significantly differ for candidates with Partisan lean data and candidates without. As a result, we proceed to build a Bayesian model to predict the number of individual contributors each candidate has, as we believe the number of individual contributors is a good indicator of support from citizens. Our model suggests that candidates in the 2018 Primary cycle have a 95% chance of raising between 29 and 945 unique donors by the election date, although we have certain reservations on the validity of its fit on the ground truth data. Although our testing shows that dropping candidates with missing Partisan Lean information does not have a significant impact on our analysis, it may still introduce a little bias because there are some inherent party leans in some of the states that were missing candidate information. There are two main limitations of this study: limitation of the dataset and limitation of methods. In our dataset, there is no information about what proposals the candidates support in their political campaign, their past experiences, and what social group they appeal to, which should play an important role in the amount of support they receive. Further research may try to gather those information about candidates' campaigns when modeling for a number of contributors. As for the Bayesian model, although it shows promising results in this study, politics is a highly complicated topic and therefore it's hard to explain its intricacies with a graphical model with a few variables. Last but not least, this study solely focuses on the Primary Election in 2018, and future research may examine other primary elections and investigate how the importance of different variables in candidate endorsement and financing have changed over time.