Predicting primary election results through network analysis of donor relationships

Michael Weingert, Ellen Sebastian

December 9, 2015

1 Introduction

In U.S. Presidential elections, campaign funding is widely regarded as an indicator of candidate success. In recent campaigns, candidates have raised upwards of 9 figures in donations from individuals and Political Action Committees (PACs). Prediction of the campaign process has been the study of economists and political scientists for centuries. However, recent efforts have taken advantage of data and analysis that has only become possible with the rising computer and Internet age.

In this project we attempt to predict the outcome of presidential primaries using network analysis of donations to campaigns. From our research, we could not find examples of this approach being applied. If successful, it could have many implications for prediction of future campaigns, as well as insightful analysis for future candidates. It can be beneficial to allow candidates to increase the efficacy of their campaigns around our analysis, and to get a meaningful metric as to how their campaign is performing thus far.

Furthermore, there have been several campaigns of late where candidates who have had significantly more funding have failed to win their respective primaries (Romney in 2008, Clinton in 2004). This paper also hopes to uncover some insights into those campaigns to help figure out why they were unsuccessful. This type of analysis will be incredibly useful and interesting to analyze the success of future campaigns. Insights will hopefully be gleaned to analyze how candidates could better fundraise to improve their odds of winning a primary.

The proposed algorithm will utilize network characteristics in a 3-stage algorithm to predict primary election outcomes: candidate drop-out, endorsements, and final nominations. Analysis is done separately for the Republican and Democratic parties. From the 2000-2012 data set, we correctly predict 8 out of 8 party nominees using PageRank, community structure, and funding measured on March 1 of the election year as features in a linear regression. Our algorithm outperforms a naive baseline which correctly predicts only 5 out of 8 nominees using only total funding information. The model predicts that Hillary Clinton and Marco Rubio will be nominated for the presidency in 2016, with Jeb Bush also competing in the final Republican party primary.

2 Previous work

2.1 Economics work

Extensive research is conducted by political analysts to attempt to gauge who will win the election. This work was first proposed by (Bean, 1948)[1] and has been expanded upon to contemporary forecasting with the work of Kramer and Tufte (1971)[2]. This work demonstrates that presidential and congressional elections coincide with political and economic conditions. (Abramowitz et al)[3] focused on developing a simple algorithm using several features: the incumbent president, condition of the economy, and timing of the election to predict the outcome of an election. Stanford Economics and Finance professor Ray Fair[4] introduced the concept of using information and data from previous elections to predict the outcome of a future election. This approach focused on a statistical prediction model to learn importance of features in prediction.

2.2 Computer Science and Machine Learning Work

Other analysis has focused on analysis of voter sentiment through twitter to predict the outcome of elections[5]. However, this analysis identified that prediction through just user sentiment has limited results. This paper wrote that social media data should not be used in isolation without accompanying research to

analyze the results of twitter sentiment analysis.

(Chappell et al) studied the effects of candidates' appearance to predict election outcomes[6, 7]. This study showed that facial appearance and perception of 'competence' was highly correlated to election results[8].

Other results showed that search volume was not a good indicator of election results[9].

Finally, (Biersack et al)[10] has shown that early donations in a campaign is a good predictor of future fund raising for members of congress and congressional campaigns. Biersack showed that money from individuals is more important than money from other sources.

2.3 Novelty of Approach

Our proposed approach builds upon the previous work outlined above. We saw through the twitter papers that popular support as measured by tweets is a valuable metric when predicting the outcome of campaigns. Furthermore, the work by (Fair et al) showed that previous elections can be used to predict future elections. His work built upon the predictors of (Abramowitz) and showed that temporal information (differences over time) is useful for prediction. Furthermore, many of the earlier papers used maching learning and linear regressions to train predictors, which we will also employ.

As a result, we incorporated the above insights, and add several novel differences outlined below:

- Treating the election process as a distinct-staged process instead of a single event. This allows us to perform more detailed analysis for each stage, instead of generalizing the entire process.
- Consider the effect of network analysis on the election. This seems to be a unique approach that differs from the literature review in that network effects (pagerank) as well as community structure are incorporated.
- Combine temporal and nontemporal information in the analysis
- Incorporate historical training sets to train for a future data set
- Aim to predict candidates for each party as opposed to overall election results.

3 Dataset

3.1 Data Collection

According to the Federal Election Campaign Act, candidates must disclose the full names, addresses, occupations, employers, and donation amounts for all donations over $200. As a result, we have access to detailed campaign finance records dating from 1980 via the the Federal Election Commission website. [1]

The relevant data on donors includes:

- donor full name, address, occupation, and employer
- donation amount (non-aggregated; multiple donations result in multiple records)
- memo text (e.g. "Payroll Deduction")
- donation recipient (a committee that forwards the donation to a candidate)
- report type, indicating at which stage in the election the donation was made; e.g. "pre-primary" or "post-convention"

The relevant data on candidates includes:

- unique candidate identifier (there is no risk of candidate duplication)
- candidate name
- office that the candidate is running for
- party affiliation
- incumbent or challenger status

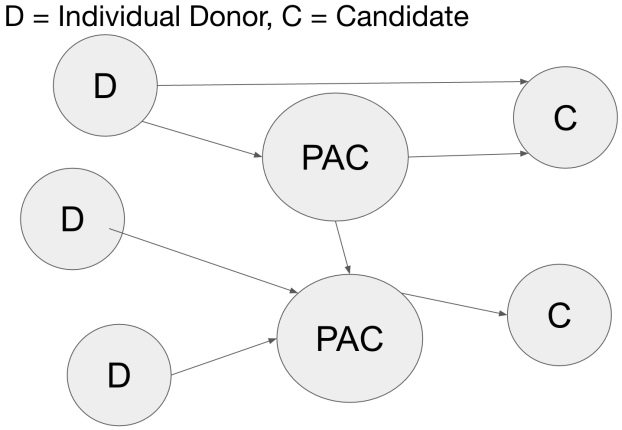
Donors and candidates are linked together via PACs. We can therefore form a pseudo-tripartite, weighted, directed graph, where edges represent donations and weights represent donation dollar amounts. The directed edges can lead from a donor to a PAC, a PAC to another PAC, a PAC to a candidate, or a donor directly to a candidate.

Summary statistics from 2008 data set:

- Average Donations per Donor = 1.3.
- Average Donations per PAC = 23.8
- Average number of donations to a candidate = 73941.

[1] http://www.fec.gov/

Figure 1: Schematic overview of donation network structure



3.2 Use of Limited Information

For elections prior to 2016, we have access to a full wealth of data from the entire campaign period. However, with the current, ongoing election we do not have access to information in the future. As a result, we trained all of our algorithm with data from before march of that election year. As the primaries are usually in June, and many candidates will begin to drop out in April and May, we chose March as a cutoff month. Therefore, all of the data in the following sections is captured from before march.

Furthermore, through our analysis we found that campaign finance information from 1980 - 2000 wildly differs from more modern campaigns (years 2000 onwards), and the completeness of data collection is suspect. The more recent usage of SuperPACs is a major change. As a result, our analysis is focused on the campaigns from 2000-2016.

It should be noted that during the period 2000-2016, two nominees were incumbent presidents: Bush in 2004 and Obama in 2012. No candidates other the incumbent were analyzed for the 2004 Republican or 2012 Democratic primary.

4 Algorithm

4.1 Baseline Algorithm

In order to analyze our results, we created a baseline algorithm. In this case, the naive baseline approach is to predict a winner to a primary based solely on funds and who has the largest amount of funds at a given time. The results of this algorithm are analyzed later.

4.2 Overview of Our Algorithm

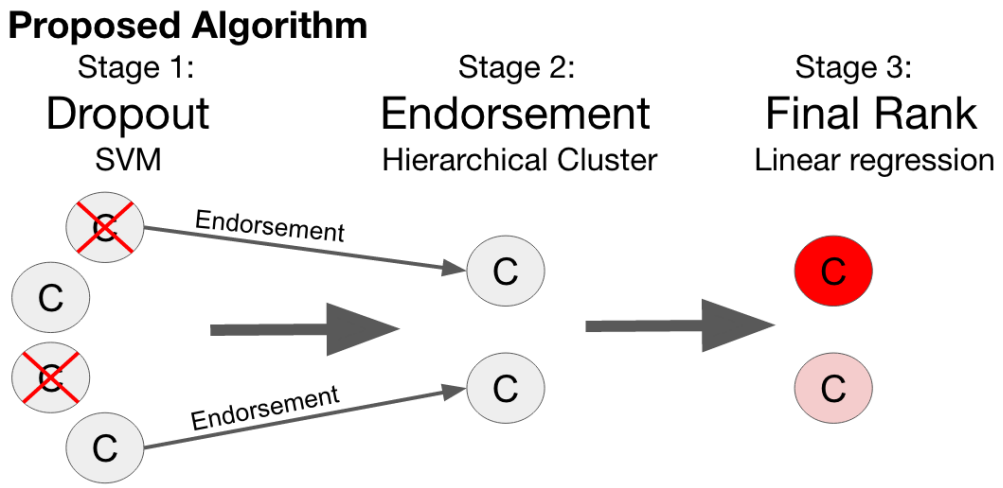
With this project we aim to predict the winner of the Republican and Democratic primaries. We hypothesize that attempting to predict the result of the overall election will be much more difficult using just donation information as there are many other factors that come into play. We hypothesize that holistic differences between parties is much harder to capture and analyze with donor information than differences between candidates within parties.

We have decided to split the party election primary process into three distinct stages. The first stage occurs when there are many candidates campaigning before the primaries. During this stage, candidates will receive donations and the candidates that are less popular (and sense that they won't win) eventually drop out.

After they drop out, a period of endorsement occurs where candidates who have dropped out 'endorse' remaining candidates. These endorsements are very important as candidates can sway their supporters to vote for their endorsed candidate. During this second stage, we will try to predict endorsements from dropped out candidates.

The final stage transitions is a vote amongst the remaining candidates. The winner of this vote amongst the remaining candidates (usually 2 or 3) determines the representative of the party who will compete against the representative of the other parties.

This algorithm is outlined in the flow chart below, where 'C' represents a candidate.



4.3 Stage 1: Dropout Prediction

The first stage is to attempt to predict which candidates will drop out. To do this, we trained a SVM on a variety of features. In order to avoid bias between candidates who dropped out vs. those who continued to the primary, the features were calculated over donations prior to March 1 on the election year.

Table 1: SVM Features Used

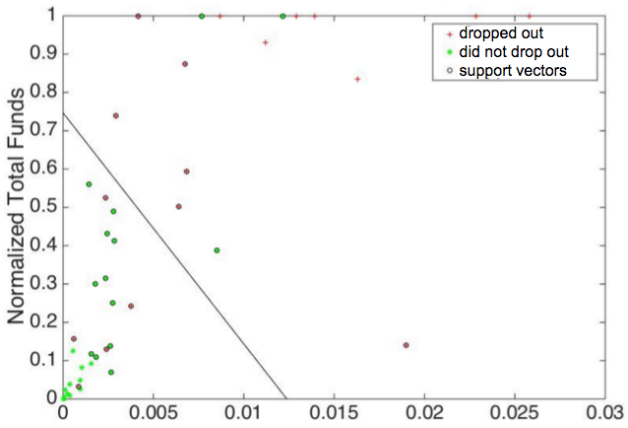
Proposed Features	Hypothesis for Importance
PageRank of Candidate	We hypothesize that a larger pagerank will correspond to a larger donor base. This means that the candidate likely has a larger support base, and is unlikely to drop out.
Total funds for candidate	We hypothesize that the donation amount going to candidates is also important as it more money allows candidates to have more advertisements, and travel for campaigning.
Funds from Individual Donors	Funding from individual donors may be beneficial to consider separately from total funding to identify Super PACs funding candidates, and see who has broader appeal.
Temporal PageRank and Temporal Weighted PageRank: PageRank / Weighted PageRank scores over time.	We hypothesize that, as candidates grow in popularity, their page ranks will also increase. Furthermore, we hypothesize that candidates who are losing pagerank are losing momentum and are more likely to drop out. The effects of temporal differences in predictions has been studied by (Sutton, 1988) . We will calculate this by looking at average slope across the last 6 months and compare that to the average slope across the first 6 months.
Louvain community size	We hypothesize that candidates' success depends not only on the number of donors to their own campaign, but also donors to other candidates whose constituency is similar to theirs. Therefore, the size of the community (number of donors, PACs, and candidates) to which the candidate belongs may be a useful feature. We noticed that the leading candidates are all contained in communities of at least size 8400, compared to the low-skewed distribution of community size and candidates per community. This trend supports the hypothesis that large community size is correlated with candidate success.
Closeness Centrality	We hypothesize that candidates are likely to endorse candidates that they are similar to. PACs and individual donors will likely endorse similar candidates to maximize the chance that one of the candidates they endorse wins. As a result, closeness centrality should be useful.

This final group of features was developed after much experimentation and tweaking with different classification algorithms as well as different feature sets. However, this combination of features produced the best results for attempting to predict the winner of the campaign primaries.

4.3.1 Training Stage 1

We trained an SVM on the features above, achieving 80% drop-out classification accuracy. Figure 2 illustrates a simplified SVM trained only on total funds and PageRank.

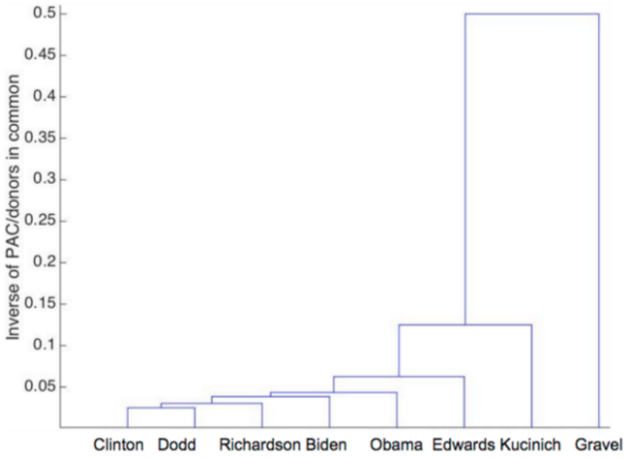
Figure 2: 2-dimensional SVM for candidate drop-out prediction. Note that if PageRank was removed, total funds would not be sufficient for an accurate classifier.



4.4 Stage 2: Endorsement prediction

The second stage is predicting, which remaining candidates dropped-out candidates will endorse. We hypothesize that having more PAC / Donors in common will be a signal of commonalities in beliefs. Thus we think that a candidate is more likely to endorse another candidate whom they have a lot of donors in common with. As a result, we performed endorsement prediction by analyzing PAC / Donors in common between candidates. Candidates will endorse the remaining candidate whom they have the most PAC / Donors in common with.

Figure 3: Hierarchical clustering reveals which candidates' supporters are similar and is used to predict candidate endorsements.



4.4.1 Stage 3: Prediction of final outcome from remaining candidates

After we have predicted endorsement, we transfer PageRank and funds from the dropped out candidate to the endorsed candidate.

The final stage is to rank the remaining candidates. A value of '0' is assigned to candidates that have dropped out, a '1' is assigned to candidates that win the primary, a '2' to candidates that are still remaining and finish second in terms of popular vote, etc.

We trained a linear regression to perform this task. Again, all features have been normalized to ensure that there is no discrepancy year to year. The normalization is done between primaries and per party.

The final list of features we used to train the algorithm is:

Table 2: Linear regression feature coefficients

Feature	Δ PageRank (Mar - Jan)	Donor funds (Mar)	Donor funds (Jan)	Louvain community size
Weight	-12.404	-0.62773	-0.44038	-0.50941

The linear regression is of the form:

$y = 1 + x1 + x2 + x3 + x4 + Intercept$, where the intercept is 3.4594.

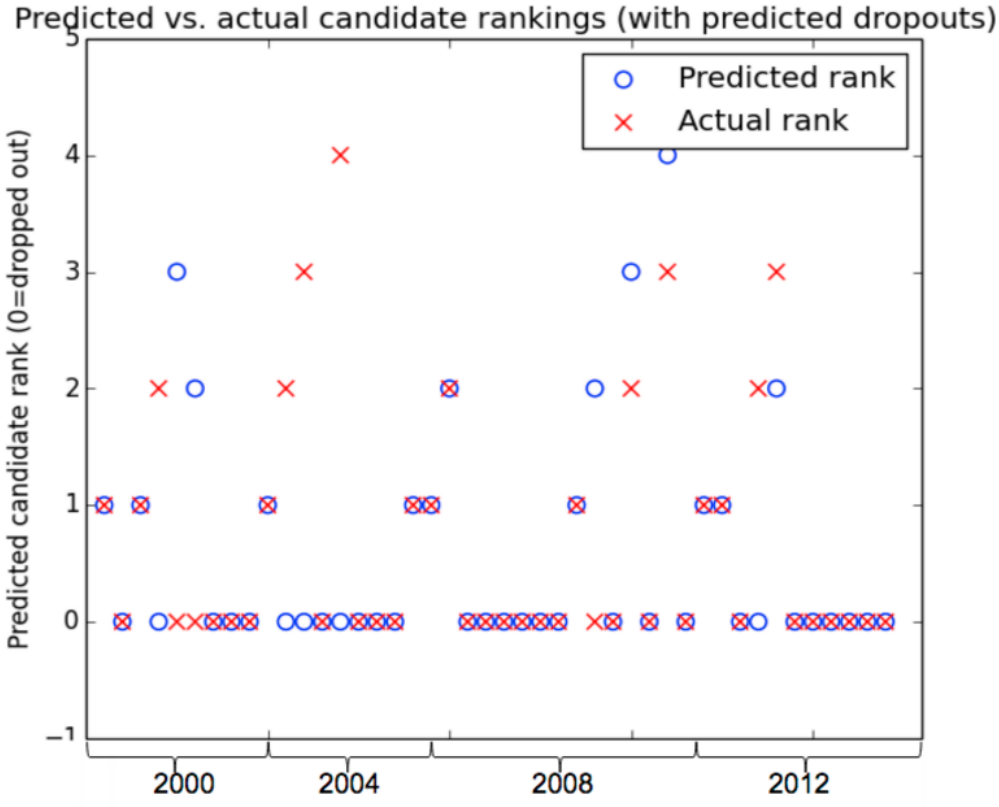
As we can see from the above analysis, the most important feature is the difference in page rank from January to March, very important months in the presidential campaign. It is also worth noting that funding from donors was more important than from PACs. Incorporating fund information from PACs actually lowered the accuracy of the prediction.

Furthermore, we notice that page rank is actually more important than funds, which supports our original hypothesis that page rank is a strong indication of the success of a candidate's campaign. These values were again discovered after much tweaking and experimentation to maximize the accuracy of the prediction.

5 Results

The graph below shows the ranking and dropout predictions produced by running all 3 stages of our algorithm.

Figure 4: Overall candidate ranking results with 100% accuracy in predicting nominee, 2000-2012



5.1 Error analysis

We can see that we correctly predict the winner of the primary in 8 out of 8 of the primaries from 2000-2012. Our accuracy is 100%. However, as we will see below, our correct prediction in the 2004 Democratic race hinged on an incorrect candidate dropout prediction; 2 mistakes canceled each other out.

We correctly predicted that 25 out of 28 dropout candidates would drop out.

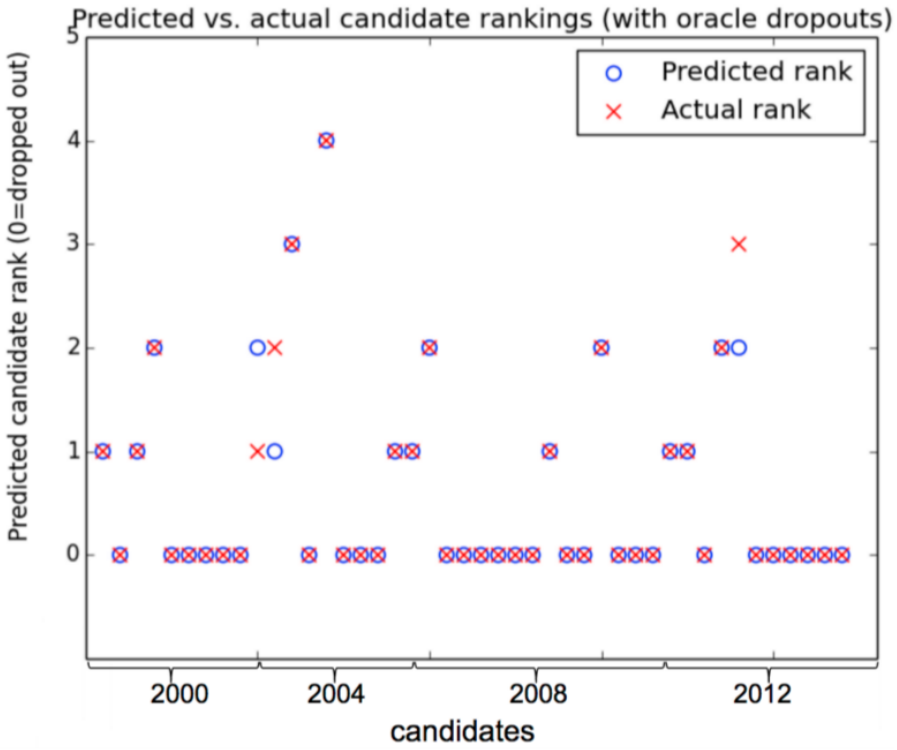
The false-negative dropouts (candidates who we predicted would not drop out, but they did) were John McCain in 2000, Mitt Romney in 2008, and Steve Forbes in 2000. The first two are notable because they would later run again and win the nomination (McCain in 2008 and Romney in 2012). Forbes' candidacy is unusual because, like Donald Trump today, he was largely self-funded[15].

The false-positive dropouts were John Edwards, Howard Dean, and Dennis Kucinich in 2004, and Newt Gingrich in 2012. It is not clear why there was an enrichment for false-positive dropouts in the 2004 Democratic primary. In 2012, Newt Gingrich had trouble fundraising and was forced to "take commercial flights and recruit professional volunteers to save money" despite his relative success in the polls; so it is not surprising that our funding-based algorithm pegged Gingrich as a dropout.

5.2 Prediction with oracle dropouts

In order to assess the impact of inaccurate dropout prediction on our final results, we also ran our candidate-ranking algorithm using "oracle" (real) candidate dropout values.

Figure 5: Candidate ranking results with oracle-provided dropouts.



Note that in the previous section, our dropout-prediction algorithm incorrectly predicted that John Edwards would drop out in 2004, thereby eliminating him as a possible nominee. However, when correct dropout predictions are given, Edwards is misidentified as the Democratic nominee for 2004 instead of John Kerry. In this case, our endorsement prediction algorithm predicted that John Edwards would be endorsed by 3 other candidates and Kerry by only one, whereas in reality Kerry was endorsed by all other Democratic candidates[16].

5.3 Comparison against baseline

In order to assess the value of network features, we decided to compare the algorithm against a naive baseline. The baseline algorithm simply ranks candidates in order of total funding, so that the predicted nominee is the candidate with the highest funding on March 1 of the election year. The predictions resulting from our algorithm and the baseline algorithm are compared with the true nominees in table 3.

In summary, our algorithm had 100% accuracy; the baseline, 62.5% accuracy; and our algorithm with oracle-provided dropouts, 87.5% accuracy.

One interesting example is the 2008 Democratic primary. In this close race, both Hillary Clinton and Barack Obama advanced to the primary with a plausible chance of nomination. Clinton actually raised more funds than Obama ($151 million vs. 126$ million on March 1, 2008). As a result, the baseline algorithm incorrectly predicted that Clinton would be nominated. However, with the aid of additional features besides total funding, our algorithm correctly identified Obama as the eventual nominee.

Table 3: Nominees, Baseline, Oracle, and Prediction, 2000-2012

Primary	Nominee	Baseline prediction	Our algorithm, oracle dropouts	Our prediction
2000 Democratic	Al Gore	Bill Bradley	Al Gore	Al Gore
2000 Republican	George W. Bush	George W. Bush	George W. Bush	George W. Bush
2004 Democratic	John Kerry	John Kerry	John Edwards	John Edwards
2004 Republican	George W. Bush	George W. Bush	George W. Bush	George W. Bush
2008 Democratic	Barack Obama	Hillary Clinton	Barack Obama	Barack Obama
2008 Republican	John McCain	Mitt Romney	John McCain	John McCain
2012 Democratic	Barack Obama	Barack Obama	Barack Obama	Barack Obama
2012 Republican	Mitt Romney	Mitt Romney	Mitt Romney	Mitt Romney
2016 Democratic	TBD	Jeb Bush	N/A	Marco Rubio
2016 Republican	TBD	Hillary Clinton	N/A	Hillary Clinton

6 Conclusion and Findings

We designed and trained a primary-election nominee predictor based on network analysis feature including PageRank and community structure. Out algorithm outperforms than a baseline that predicts candidates using funding statistics only, showing that it is capable of predicting candidate success beyond the surface of summary fundraising numbers. We have learned several interesting takeaways from this experiment.

PageRank is a better indicator of success than Funds From the final rank prediction algorithm (Stage 3), we have discovered that PageRank is a better indicator of success than amount of funds alone. This insight allowed us to train a better prediction algorithm than the baseline.

Drop out prediction is Difficult Candidates can be very unpredictable when it comes to dropping out. Some candidates who are clearly losing will decide to not drop out for a variety of reasons. Conversely, some candidates who are polling much better will realize that they don't have a strong chance of winning and will drop out. As a result, it becomes difficult to predict if or when candidates will drop out.

Predicting the winner from remaining candidates is Easier From our analysis with a drop out oracle, we were able to predict, with a high amount of accuracy, which candidates will win the election, and what the ordering will be of the remaining candidates.

Funds from individual donors is a better predictor of success than funding from PACs This analysis showed that, even though candidates may have more total funding, the amount of funding from individual donors is a better predictor of success than total funding.

Network Analysis of Donations can be used to Predict Party Primaries With our algorithm, we were able to accurately predict the outcome of all of the primaries from 2000-2012. This shows that donations are in fact a useful metric to consider when analyzing political campaigns.

Limitations. Our algorithm is not well-equipped to deal with candidates who are self-funded, such as Steve Forbes in 2000 and Donald Trump in 2016. And of course, the algorithm does not take into consideration any candidate success factors other than funding. To get a clearer picture of candidates' position from a non-financial stanpoint, future network analysis work could investigate other indicators of candidate popularity, such as online discussion, mailing list engagement, and event attendance.

6.1 Distribution of Work

Report writing was split evenly between Ellen and Michael. In addition, the individual contributions are outlined below.

6.1.1 Ellen

Ellen calculated the community structure, weighted PageRank, and n-clique overlap features in python, and was responsible for initial data collection.

6.1.2 Michael

Michael was responsible for the machine learning in Matlab for the 3 stages in the algorithm (feature selection, machine learning algorithm selection, experimentation, and tweaking). Michael was also responsible for collection of several large features in c++ including PageRank, funds (from donors, and PACs), and counting of PACs/Donors in common.

References

[1] Bean, Louis. 1948. How to predict elections. New York: Knopf.

[2] Kramer, Gerald H. 1971. Short-term fluctuations in U.S. voting behavior. American Political Science Review 65:131-43

[3] Abramowitz, Alan. "An Improved Model for Predicting Presidential Election Outcomes." Political Science and Politics 21.4 (1988): 843-47. JSTOR. Web. 2015.

[4] Fair, Ray. Predicting Presidential Elections and Other Things, Second Edition. Palo Alto: Stanford UP, 2011. Web.http://fairmodel.econ.yale.edu/rayfair/pdf/2011b.pdf

[5] Andranik, Tumasjan. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. Web.

[6] Ballew, C. C., and A. Todorov. "Predicting Political Elections from Rapid and Unreflective Face Judgments." Proceedings of the National Academy of Sciences(2007): 17948-7953. Web. http://www.pnas.org/content/104/46/17948.short

[7] Lawson, Chappell, Gabriel S. Lenz, Andy Baker, and Michael Myers. "Looking Like a Winner: Candidate Appearance and Electoral Success in New Democracies." World Pol. World Politics: 561-93. Web. http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=7909925&fileId=S0043887110000195

[8] Todorov, A. "Inferences of Competence from Faces Predict Election Outcomes."Science (2005): 1623-626. Web. http://www.sciencemag.org/content/308/5728/1623.short

[9] Catherine Lui, Panagiotis T. Metaxas and Eni Mustafaraj (2011). "On the Predictability of the U.S. Elections through Search Volume Activity," Proceedings of the IADIS International Conference on e-Society, Avila, Spain, March, 2011.http://repository.wellesley.edu/scholarship/23/

[10] Biersack, Robert, Paul S. Herrnson, and Clyde Wilcox. "Seeds for Success: Early Money in Congressional Elections". Legislative Studies Quarterly 18.4 (1993): 535–551. Web. http://www.jstor.org/stable/439854

[11] Sutton, Richard S. "Learning to Predict by the Methods of Temporal Differences."Mach Learn Machine Learning: 9-44. Web. http://link.springer.com/article/10.1007/BF00115009

[12] Adamic, Lada A., and Natalie Glance. "The Political Blogosphere and the 2004 U.S. Election." Proceedings of the 3rd International Workshop on Link Discovery - LinkKDD '05. Web. http://dl.acm.org/citation.cfm?id=1134271.1134277

[13] Brown, Lloyd. "Forecasting Presidential Elections Using History and Polls."International Journal of Forecasting 15.2 (1992): 127-35. Web. http://www.sciencedirect.com/science/article/pii/S0169207098000594

[14] Granka, Laura. "Using Online Search Traffic to Predict US Presidential Elections." PS: Political Science And Politics APSC (2013): 271-79. Web

[15] Weber, Doug. 2011. "Unlimited Presidential Fundraising: The Curse Of Steve Forbes". OpenSecrets.org. http://www.opensecrets.org/news/2011/12/unlimited-presidential-fundraising/

[16] Wikipedia contributors. "John Kerry presidential campaign, 2004." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 3 Oct. 2015. Web. 9 Dec. 2015.

7 Appendix: Summary of all features collected

This table contains all the network and non-network features we have collected about the graph for 2008 Democratic Presidential candidates. We collected this data about all major candidates from presidential elections between 1980 and 2016. The full dataset can be found here.

Candidate	Obama	Clinton	Edwards	Biden	Dodd	Gravel	Kucinich	Richardson
Total donation value ($)	72,713,998	106,423,137	32,988,769	11,901,618	15,729,249	184,663	1,127,750	1,895,4294
Number Direct donors	342203	126768	22002	6407	6197	310	1879	14831
Number indirect donors (via PACs)	455404	235171	56818	20848	43153	1540	6328	69570
Number PACs contributing to candidate	120	430	22	76	229	6	19	101
PageRank	0.0113	0.0158	0.00331	0.00204	0.003555	3.76E-05	0.000433	0.00440
Temporal PageRank	-2.636e-04	0.003	-0.001	-0.001	4.639e-04	1.868e-05	2.449e-04	9.486e-04
Weighted PageRank	0.018	0.014	0.0041	0.0013	0.0014	7.00e-05	0.00044	0.008
Temporal Weighted PageRank	-9.557e-04	-6.144e-04	-0.001	-3.059e-05	-6.230e-05	-9.801e-07	2.449e-04	3.726e-04
Closeness Centrality	0.315	0.343	0.308	0.313	0.321	0.245	0.276	0.324
average n-Clique overlap (all donors)	0.0363	0.0507	0.0576	0.0893	0.0743	0.051	0.0838	0.0684
average n-Clique overlap (PACs only)	0.173	0.227	0.2459	0.3087	0.2458	0.0501	0.1857	0.2845
Louvain community size	53631	46478	54645	54645	54645	8893	8400	15758
Num. Candidates in Louvain community	7	2	14	26	23	9	14	17
Num. PACs in Louvain community	14	11	55	97	79	37	29	48

PAC / Donor Overlap between Candidates

Candidate	Obama	Clinton	Edwards	Biden	Dodd	Gravel	Kucinich	Richardson
Obama	0	23	11	9	11	2	6	13
Clinton	23	0	16	26	40	2	8	33
Edwards	11	16	0	9	9	2	7	11
Biden	9	26	40	0	21	2	7	20
Dodd	11	40	9	21	0	2	7	22
Gravel	2	2	2	2	2	0	2	2
Kucinich	6	8	7	7	7	2	0	8
Richardson	13	33	11	20	22	2	8	0