Elections & Opinion Polls

Marcus Sinclair

May 23, 2022

Contents

1 Objectives **2**

2 Methodology & Outcome **2**

3 Data Understanding & Preparation **3**

3.1 Categorical Data . . . 3

3.1.1 Bar Plot of Opinion Polling Company Survey Frequency . . . 3

3.1.2 Bar Plot of Opinion Polling Company Total Sample Size . . . 4

3.2 Numerical Data . . . 5

3.2.1 Scatter Plot of Opinion Polling Company Results w.r.t Time . . . 5

3.2.2 Scatter Plot of Opinion Polling Company Results & Uncertainties . . . 6

4 The 2016 US Election Instability **7**

4.1 Non-Response Bias & Deficient Weighting . . . 7

4.2 Misspecified Likely Voter Models . . . 7

4.3 Late Deciding & The “Fake News” Epidemic . . . 8

4.4 The Shy Trump Hypothesis . . . 9

5 Evaluation **9**

5.0.1 Table Quantifying Late Deciding Hypothesis In 2016 US Election . . . 10

6 Conclusion **11**

1 Objectives

Pre-election opinion polling is the process of surveying a random sample population of a ward or state in order to determine the preferred majority party, thus in essence attempt to predict the election outcome. Using a data analytical approach, our goal is to **investigate how opinion polling companies try to estimate voting intentions**. We shall focus specifically on pre-election opinion polls for the 2016 US election with the aim to discuss:

1. **How do companies differ in their practice and their accuracy?**
2. **How stable are opinion polls, and what are the typical uncertainties involved?**

The dataset used is a subset of the [Nationwide opinion polling for the 2016 United States presidential election data source] - table 2. The dataset contains results from 39 unique opinion polling companies, 7 variables spanning 264 separate surveys. Before we wrangle and explore opinion polling company data, we must gain some contextual understanding of the US electoral process.

2 Methodology & Outcome

Analogous to the UK Parliament, the United States Congress has two houses, the House of Representatives and the Senate. The main differences arises due to the fact that there are separate elections for both houses and a third separate election to determine the presidency electorate. In the Electoral College system, each state gets a certain number of electors based on its total number of representatives in Congress. Each elector casts one electoral vote following the general election; there are a total of 538 electoral votes. The candidate that gets more than half (270) wins the election.
Donald Trump won the general election on Tuesday, November 8, 2016, despite losing the popular vote while winning the electoral college. [1] Most polls correctly predicted a popular vote victory for Clinton, but overestimated the size of her lead, with the result that Trump's electoral college victory being a surprise to analysts. With this observation in mind, let us explore outputs of the 39 opinion polling companies for the 2016 US election.

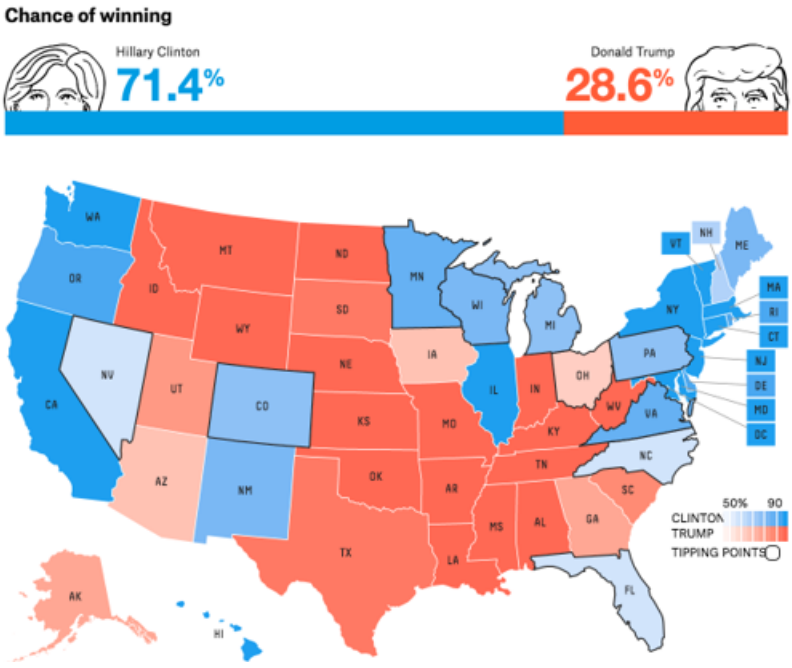

Figure 1: The prediction for the 2016 US election by Nate Silver's FiveThirtyEight website.

3 Data Understanding & Preparation

3.1 Categorical Data

3.1.1 Bar Plot of Opinion Polling Company Survey Frequency

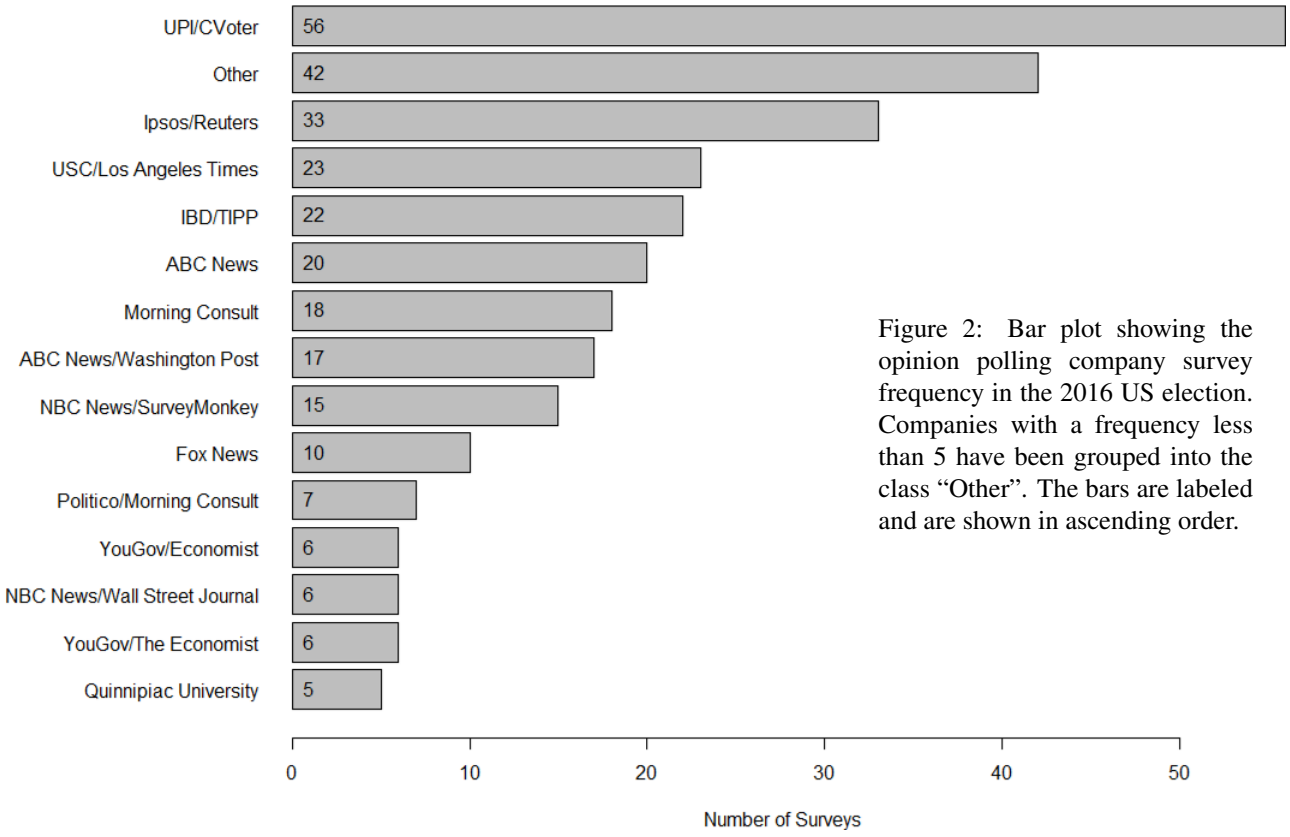

Figure 2: Bar plot showing the opinion polling company survey frequency in the 2016 US election. Companies with a frequency less than 5 have been grouped into the class "Other". The bars are labeled and are shown in ascending order.

Producing more than a fifth of total surveys issued, UPI/CVoter produced the most frequent amount of polls - 56, almost double the second most frequent company that is, Ipsos/Reuters with a survey total of 33. Their methodology focuses on breadth, surveying 73,342 US residents on their opinion about the election.

CVoter uses online opinion polls. Statistical margins of error are not applicable to online polls [10]. The precision of online polls is measured using a credibility interval. The error due to sampling for projections based on the Likely Voter sample could be plus or minus 3 percentage points at the national level and plus or minus 5 percentage points at state level. All sample surveys and polls may be subject to other sources of error, including but not limited to coverage error- not a 1-1 correspondence between the target population and the sampling frame, and measurement error- the difference between a measured quantity and its true value i.e. the sampling frame provide false opinions.

In the State Tracker, CVoter employ multiple providers of panels to randomize and remove the contact bias of any one particular sample provider. CVoter use an exclusive "Psephometer" algorithm which was updated on a daily basis during election season. This means they have the national projection as well as the state-level projections on a daily basis - this is a unique tracker relative to other opinion poll companies from that perspective.

3.1.2 Bar Plot of Opinion Polling Company Total Sample Size

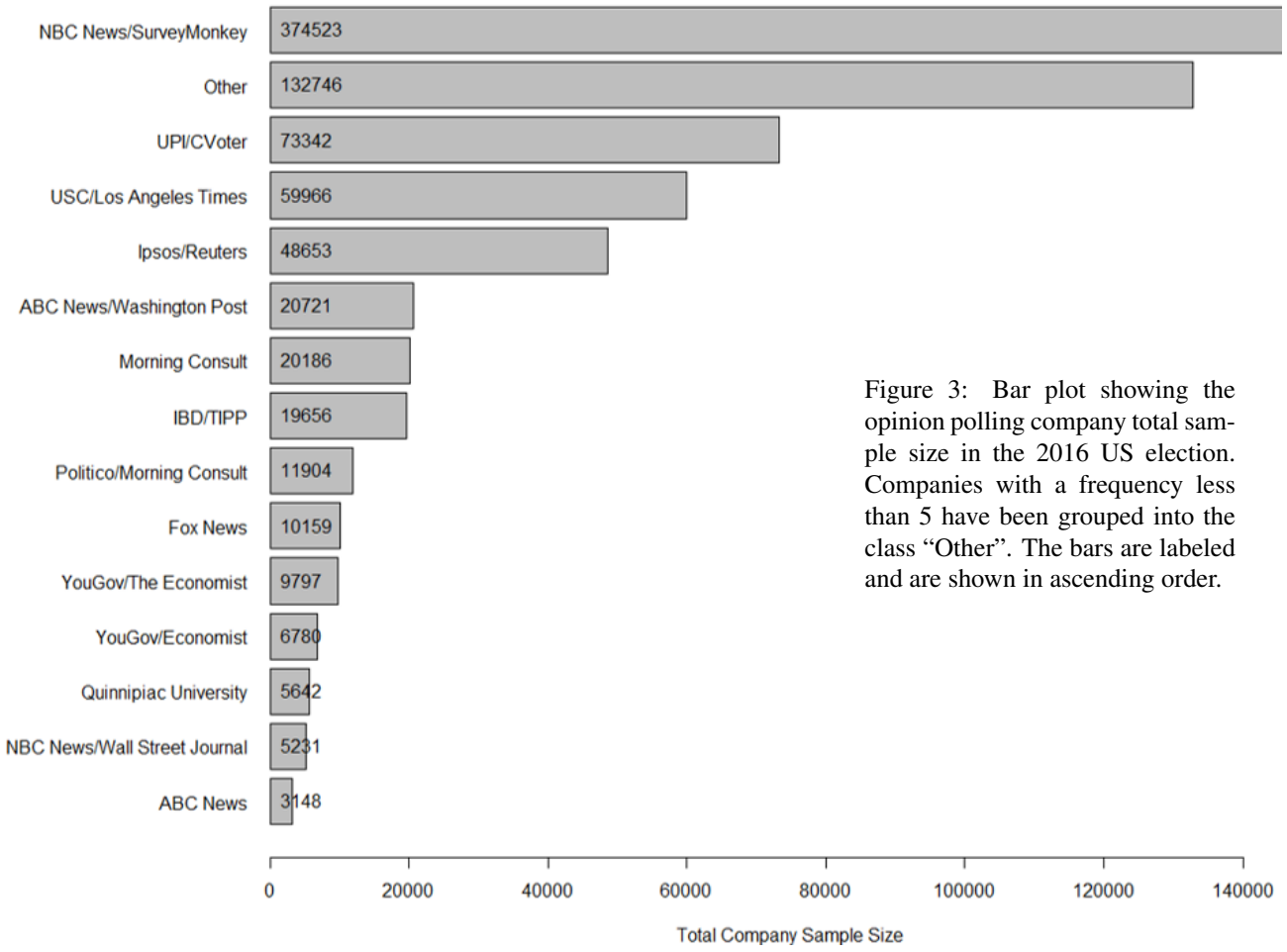

Figure 3: Bar plot showing the opinion polling company total sample size in the 2016 US election. Companies with a frequency less than 5 have been grouped into the class "Other". The bars are labeled and are shown in ascending order.

As seen in figure (3), the vast majority of total US residents polled were acquired from the NBC News/SurveyMonkey framework, surveying 374,523 residents. Similar to CVoter, SurveyMonkey conducts public opinion polls from the pool of users of its other online surveys. The company asks a random portion of users to participate in additional surveys and then adjusts the results to be representative of the population [6]. Regarding the 2016 election, the Washington Post wrote, "Altogether, our review found SurveyMonkey estimates to be broadly in line with election results, other polling benchmarks and our own trusted cellular and landline phone surveys. These comparisons gave us confidence that results from the sample can be useful in shedding light on the opinions on voters" [4].

The USC/Los Angeles Times opinion poll spanned a reasonable 59,966 residents. Regarding their methodology, interviews were conducted by telephone using live interviewers from "Interviewing Services of America" [2]. Voters were randomly selected from a list of registered voters statewide and reached on a landline or cell phone depending on the number they designated on their voter registration. 55% of this sample was reached on a cell phone. Up to 5 attempts were made to reach and interview each randomly selected voter. The data was weighted to reflect the total population of registered voters throughout the state, balancing on regional and demographic characteristics for gender, age, race, and party registration according to known census estimates and voter file projections from several distinct voter files.

3.2 Numerical Data

3.2.1 Scatter Plot of Opinion Polling Company Results w.r.t Time

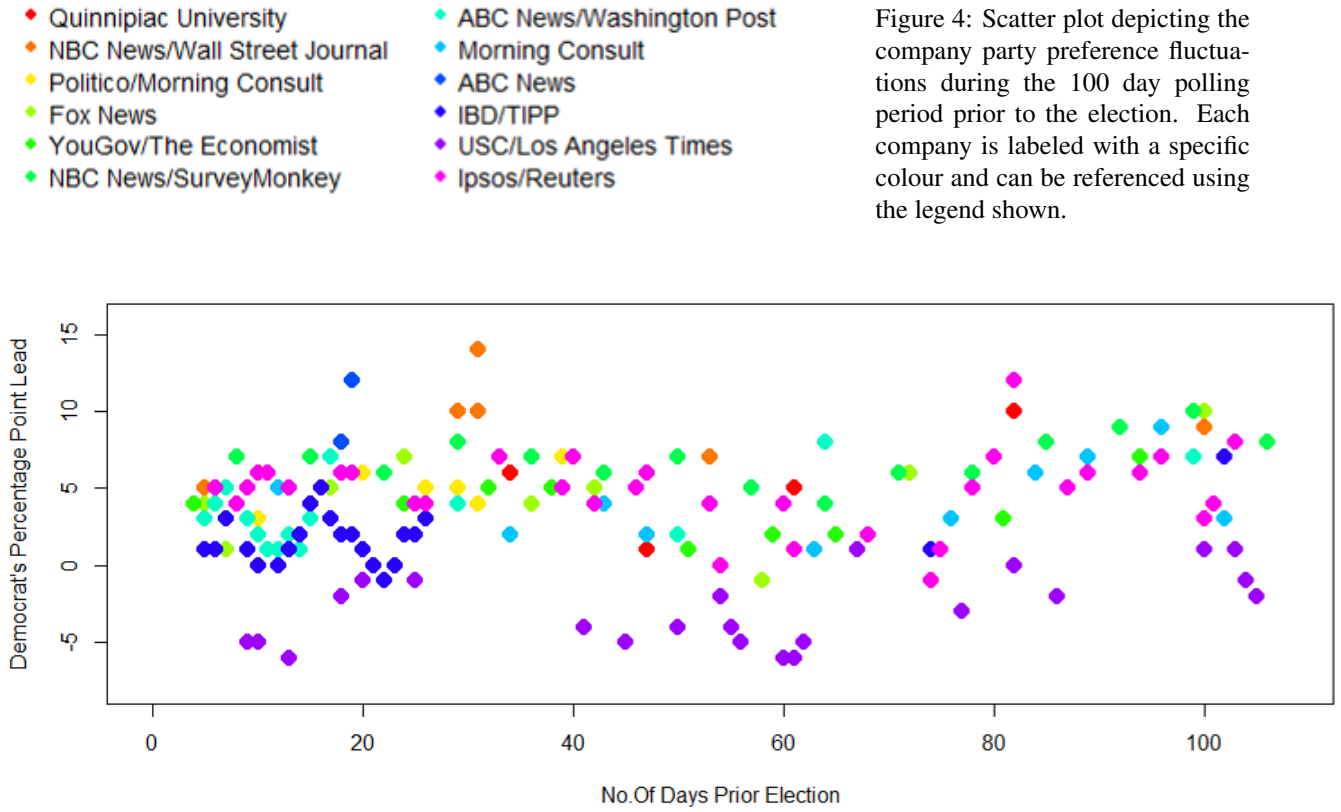

Figure 4: Scatter plot depicting the company party preference fluctuations during the 100 day polling period prior to the election. Each company is labeled with a specific colour and can be referenced using the legend shown.

Many inferences can be made from figure (4). Firstly, instability in polling outcomes is shown with companies outputting large deviations in Democrat's percentage point lead during the pre-election phase throughout their surveys. A negative Democrat percentage indicates a Republican vote majority. For instance, USC/ Los Angeles Times obtains a maximum output of 0 (neutral party preference) and a minimum of -6 (large Republican preference) throughout the pre-election process, with a standard deviation of 2.36 throughout their 23 surveys issued.

Furthermore, differences in company practices can be seen. IBD/TIPP produced daily surveys for 3 weeks prior to the election and only 2 other surveys spanning the remaining pre-election time period. In contrast to this methodology, YouGov/The Economist produced 21 surveys, equally spread out over the 100 days.

Excluding the USC/ Los Angeles Times output, polling outputs are roughly correlated within the time frame. For example, 40 days prior to the election, we see a general downturn towards preferences becoming more Republican favoured as seen by the general decrease in Democrats percentage point lead for the opinion polling companies. Similar polling swings are desirable to see as daily changes in Democratic and Republican party policies should sway resident votes in a somewhat similar fashion throughout the prediction polls.

3.2.2 Scatter Plot of Opinion Polling Company Results & Uncertainties

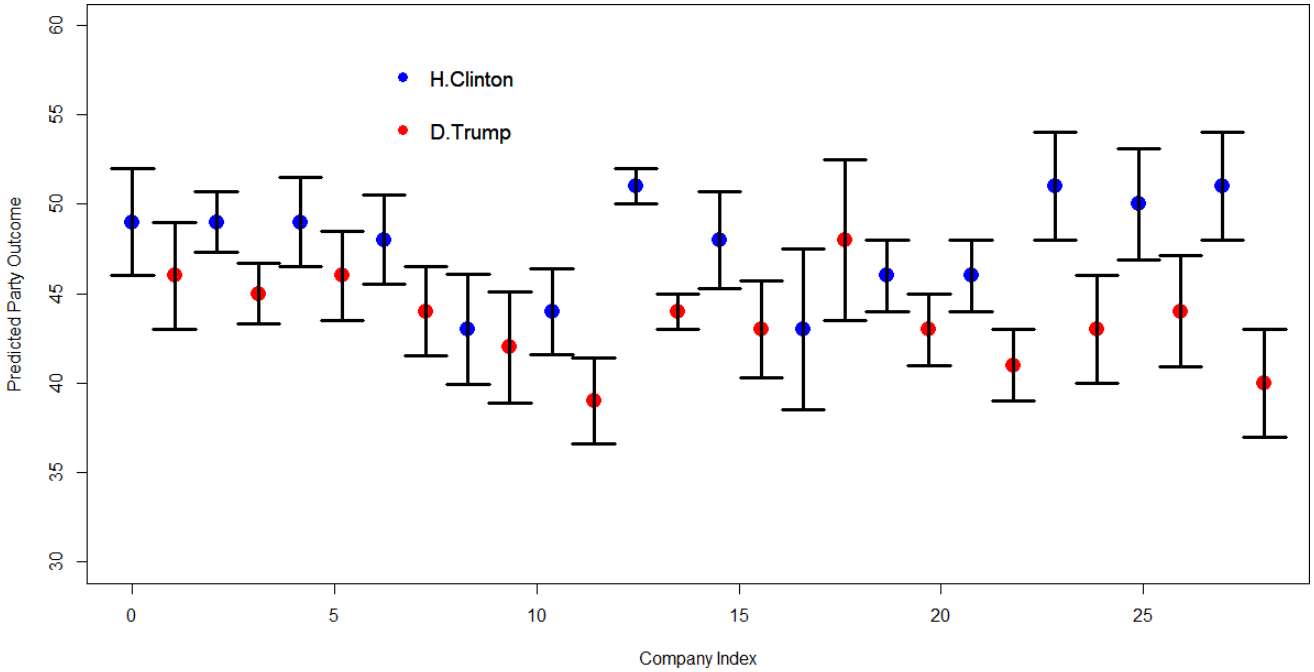

Figure 5: Scatter plot of opinion polling company results. The 15 of the 39 opinion polling companies with the largest sample sizes were used and are referenced as an index. Results were taken from the polls closest to election day. Blue dots indicate predictions for Hillary and red for Donald. The whiskers show margins of error as stated by the respective opinion polling company.

As seen in figure (5), Hillary Clinton is the electoral favorite for almost all company polls, with the only exception coming from the USC/ Los Angeles Times, predicting a H:D (Hilary:Donald) output of 43%:48%. That being said, the margins of error in opinion polling outcomes span a large predicted party output for the vast majority of companies. For instance, IBD/TIPP polled a H:D output of 43%:42% with a margin of error of 3.1%. Companies output unstable results and thus their predictions must be interpreted with care.

Reasons for instability could be due to different types of polls (e.g., online versus live telephone) seemingly producing somewhat different estimates. This question became the central foci for an ad hoc committee commissioned by the **A**merican **A**ssociation for **P**ublic **O**pinion **R**esearch (**AAPOR**) in the spring of 2016 [8]. The committee was tasked with summarizing the accuracy of 2016 pre-election polling, reviewing variation by different poll methodologies, and assessing performance through a historical lens. Their response spanned the topics of:

1. **Non-response bias and deficient weighting**
2. **Late deciding & the "Fake News" epidemic**
3. **Misspecified likely voter models**
4. **The shy trump hypothesis**

4 The 2016 US Election Instability

4.1 Non-Response Bias & Deficient Weighting

Non-response bias occurs when non-responders from a sample differ in a meaningful way to responders. This bias is common in descriptive, analytic and experimental research and it has been demonstrated to be a serious concern in survey studies. In the context of the 2016 US election, one could argue that the general characteristics and traits of a strong Republican American resident results in a skewness of pre-election survey results. Republicans are less likely to participate in opinion polled company surveys in comparison to other party members. [8] Given the anti-elite themes of the Trump campaign, Trump voters may have been less likely than other voters to accept survey requests. If survey response was correlated with presidential vote and some factor not accounted for in the weighting, then a deficient weighting protocol could be one explanation for the polling errors.

As seen in figure (4), an opinion polling company going against the pro-Democratic trend is the USC/ Los Angeles Times. Throughout the pre-election period, the surveys issued by the USC did not favour Hillary like the vast amount of the other 38 polling companies did; USC favoured Trump, growing in preference up to a 6 point lead in the Republican campaign 5 days prior to the election. [9] As Ernie Tedeschi, a Washington-based economist and former Treasury Department official, has shown, if you take the Daybreak poll's data — which USC made available to the public — and weight it more in line with the usual system pollsters use, you get results that largely match the polling averages. In this extreme example, instead of the general issue of a deficient weighting in Republican opinions, one could argue that USC over exaggerated Trump supported views, resulting in a preference bias not towards the Democratic party but the Republican party instead.

4.2 Misspecified Likely Voter Models

As mentioned during discussions of USC and SurveyMonkey methodology, polling companies must construct a likely voter model. Companies attempt to predict which Americans are most likely to vote in the 2016 election, matching their pre-election opinion polling surveys accordingly. A popular procedure within these companies is to use past-data from previous elections, extrapolating the information gathered during these previous elections to the present day.

A possible issue with this method can be seen if we consider how the likely voter model data is collected historically. Gallup, one of the 39 polling companies, asks poll respondents a variety of questions about their interest in the coming election, their past voting behavior, and their intention to vote in the coming election. [7] For the 7 questions that make up the likely voter scale, respondents receive 1 point on the likely voter scale for each question to which they give the response listed in parentheses (with a maximum of 7 points possible).

1. Thought given to election (quite a lot, some)
2. Know where people in neighborhood go to vote (yes)
3. Voted in election precinct before (yes)
4. How often vote (always, nearly always)
5. Plan to vote in 2012 election (yes)
6. Likelihood of voting on a 10-point scale (7-10)
7. Voted in last presidential election (yes)

This is a common likely voter methodology for opinion polling companies. The issue of the use of historic data collected in this fashion is clearly evident in 2016, for example, voter turnout in 2016 differed from that in 2012 in ways that advantaged Trump and disadvantaged Clinton. [8] Nationally, turnout among African Americans, the group most supportive of Clinton, dropped 7 percentage points while turnout among Hispanics and non-Hispanic whites changed little. If pollsters designed their likely voter models around the assumption that 2016 turnout patterns would be similar to 2012, this could have led to underestimation of support for Republicans, including Trump.

4.3 Late Deciding & The "Fake News" Epidemic

With the rise of social media in recent times, the plasticity of a US resident's vote has been more pronounced than ever. The strong Republican views are more likely to be accepted if one sees a community of similar viewed minds on the matter. Twitter could be argued to have given a platform for strong right-wing viewpoints to be shared freely throughout the country, molding residents party preferences during the pre-election period. Trump himself declared after becoming president: "I think I wouldn't be here if I didn't have social media".

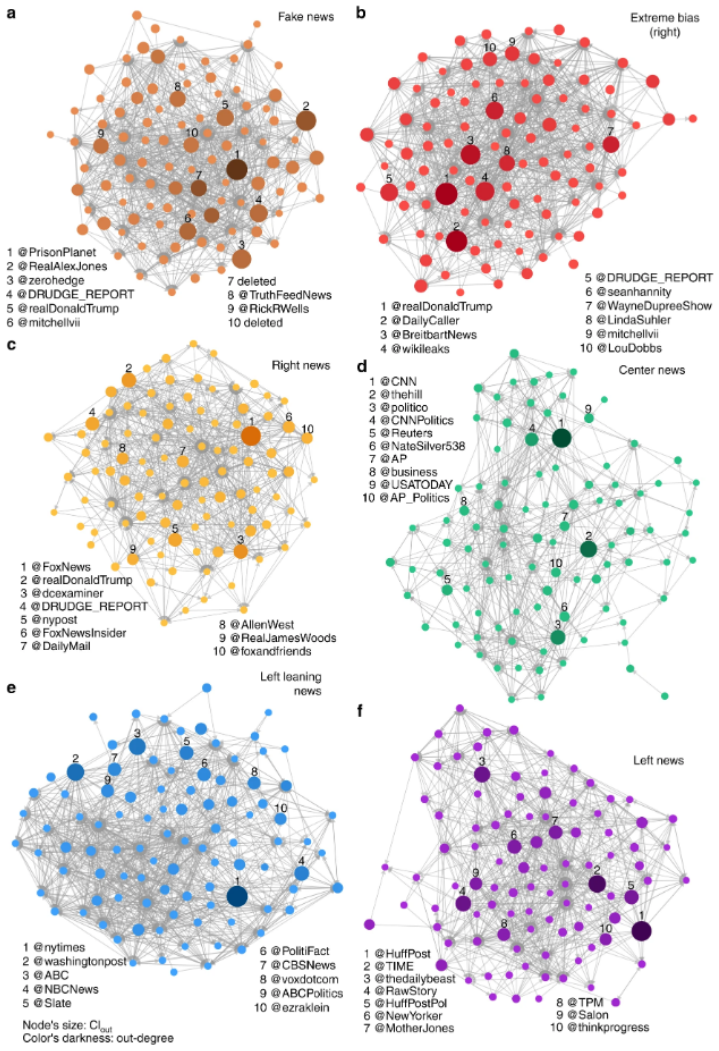

Figure 6: Retweet networks formed by the top 100 news spreaders of different media categories produced by Bovet, A. The direction of the links represents the flow of information between users. The size of the nodes is proportional to their Collective Influence score, and the shade of the nodes' color represents the number of different users that have retweeted at least one of her/his tweets with a URL directing to a news outlet, from dark (high number) to light (low number). The network of fake (a) and extreme bias (right) (b) are characterized by a connectivity that is larger in average and less heterogeneous than for networks of center and left leaning news

Fake news that is, news stories that are false, fabricated, with no verifiable facts, sources or quotes was more prevalent than ever in the 2016 US election. To gain insight in the scale of the fake news epidemic in the election, Bovet, A. a researcher at the Levich Institute and Physics Department New York, analysed 30 million tweets, from 2.2 million users, which contain a link to news outlets during the pre-election time frame. He found that 25% of these tweets spread either fake or extremely biased news, with the activity of Trump supporters influencing the dynamics of the top fake news spreaders [3].

Figure (6) depicts the hive-mind twitter structure on the 2016 US election represented in 6 subsets. Retweet networks for fake news (a), extreme bias (right) news (b), right news (c), center news (d), left leaning news (e), and left news (f) showing only the top 100 news spreaders ranked according to their collective influence. The structural differences shown in the networks (a) and (b) may be explained by the fact that there is something different about the way that the people in these networks organize and share information. Bovet suggests that while this may be true, it may also be the case that there are subgroups of users in the center and left leaning news networks that form diffusion networks with a similar structure as the smaller fake and extremely biased news networks and then also have a large number of other individuals added to these subgroups due to the presence of important broadcast networks that feed their ideology or information needs.

4.4 The Shy Trump Hypothesis

Yes, tools such as social media have been used to unite marginalized communities and find like-minded individuals as mentioned prior however, there is evidence to suggest opinion polling reporting error may still partially arise from under-reporting residents backing controversial candidates. Some Trump voters may not have been willing to disclose their support for him in surveys [8].

Albeit similar to the late deciding phenomenon, instead of residents changing viewpoints during the final few weeks prior to the election, the shy trump hypothesis claims residents with controversial party preferences not only lie about which party they support, but stay away from opinion polls altogether, giving bias to opinion polling data. There are many reasons for a US resident to feel pressured in this fashion during the 2016 US election. Clinton was the first female major-party presidential nominee, and although both candidates were white, Trump's record on racially charged issues and open support from white supremacists put race in the forefront of the campaign.

This results in disparities seen in figure (7) whereby due to the controversial candidates present in 2016, a large vote margin is present of 16% towards Republicans relative to other years seen in the chart - inconsistencies in polling and post election outcomes are present.

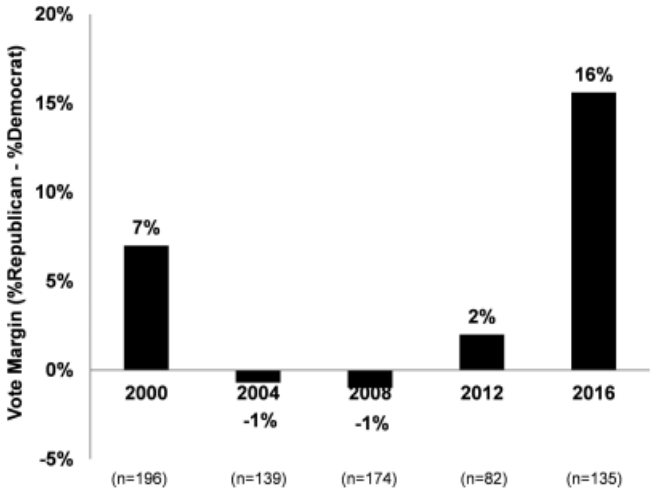

Figure 7: Bar chart produced by Kennedy et al showing the vote margin (% voted for Republican candidate – % voted for Democratic candidate) among callback respondents giving inconsistent pre- vs. postelection responses. Data are from Pew Research Center RDD callback studies.

5 Evaluation

Given a dataset containing results from 39 unique opinion polling companies on the 2016 US election, 7 variables spanning 264 separate surveys, we have wrangled and explored the data, described company differences in methodology and observed instability in opinion polling outcomes. As shown in section (3.1), online surveys and questionnaires are at the forefront of company methodology, with the accuracy and precision of these surveys diminishing mainly due to measurement and coverage error. As mentioned prior, online survey error could be plus or minus 3 percentage points at the national level and plus or minus 5 percentage points at state level. Despite these large inaccuracies present, online opinion polling will remain the most popular choice for polling companies to implement. There are numerous reason for this. Choices optimal from the point of view of precision such as paper polling, quotas, and in-person questionnaires are costly in money and time [5]. It should scarcely be concluded that "total error" in survey research can be reduced to mere sampling error simply by "throwing money at the problem". Opinion polling companies apply cost-benefit analysis in regards to their potential method and resources - a balance in expense and accuracy must be made.

Following our discussion of company inaccuracy, we explored the stability of opinion polling outcomes, showing in section (3.2) large deviations in polling outcomes during the 2016 US pre-election phase and overlapping uncertainties in Democratic vs Republican preferences throughout the 39 companies surveying in this time frame. Reasons considered to produce these disparities present were:

1. **Non-response bias and deficient weighting**
2. **Misspecified likely voter models**
3. **Late deciding & the "Fake News" epidemic**
4. **The shy trump hypothesis**

For the Non-response bias and deficient weighting argument, whilst one could argue Republicans are less likely to participate in opinion polled company surveys in comparison to other party members, exceptions to this line of reasoning can be found. USC did not favour Hillary like the vast amount of the other 38 polling companies did; USC favoured Trump, growing in preference up to a 6 point lead in the Republican campaign 5 days prior to the election. Careful attention to the potential problem of non-response bias is a critical step in conducting high quality research using survey data. Polling companies seen in the US election did indeed attempt to take into account this discrepancy in their algorithms. For instance, CVoter's "Psephometer" algorithm weighted for contact and non-response bias [10]. The level of measurability of non-response bias could be questioned. Further analysis into the weighting of such opinion polling algorithms should be made to determine whether some level of deficiency is present.

From a bias at the voters end, opinion polling instability may also arise from misspecified likely voter models as discussed in section (4.2). To evaluate the extent to which this phenomenon is most prevalent, an approach is to validate which respondents voted and which did not. This process will take into account how accurately the pollster's methods identified likely voters, and on whether either non-voters included in the sample or actual voters left out contributed to any error in estimating the ultimate result. In practice however, a full validation is neither easy nor feasible for the vast majority of public polls seen in the US election [8]. Polls conducted by telephone rarely attempt to ask and record the full name and street address of every respondent and online questionnaires do not span a desired amount of user information that would be required to go forth with rigorous analysis. For example, Kennedy et al post-election analysis determined that whilst SurveyMonkey polls conducted online were clearly improved in accuracy via non-response bias weighting, their likely voter model tended to have little to no effect on SurveyMonkey's estimates for the states examined.

5.0.1 Table Quantifying Late Deciding Hypothesis In 2016 US Election

	% Voters who decided in final week	Vote choice among voters deciding in final week		Vote choice among voters deciding earlier		Estimated Trump gain from late deciders	Election (% Trump – % Clinton)
		Trump	Clinton	Trump	Clinton		
Florida	11%	55%	38%	48%	49%	2.0%	1.2%
Michigan	13%	50%	39%	48%	48%	1.4%	0.2%
Pennsylvania	15%	54%	37%	50%	48%	2.3%	1.2%
Wisconsin	14%	59%	30%	47%	49%	4.3%	0.8%
National	13%	45%	42%	46%	49%	0.8%	–2.1%

Table 1: Table produced by Kennedy et al showing the percentage of voters who decided in the final week prior to the 2016 US election across 4 states and the national average. Late deciding voters only improved Trump's ratings across all platforms.

As seen in table (1), the late deciding hypothesis was prevalent in 2016, swaying substantial amount of US residents towards a Republican party preference in the final week of the campaign. Kennedy et al deemed the four states Hilary lost with the smallest margins to Donald to be in majority determined by late deciding voters. That is Michigan, Wisconsin, Pennsylvania, and Florida, 11 to 15 percent of voters said that they finally decided for whom to vote in the presidential election in the last week [8]. One could argue that the sheer volume of fake news circulating social media during this time frame accounted for a large proportion of this sway in party preferences seen. According to the exit poll, these voters broke for Trump by nearly 30 points in Wisconsin, by 17 points in Pennsylvania and Florida, and by 11 points in Michigan. Upon analysis, Kennedy remarks that "If late deciders had split evenly in these states, the exit poll data suggest that Clinton may have won both Florida and Wisconsin, although probably not Michigan or Pennsylvania, where Trump either won or tied among those deciding before the final week".

Finally, the shy Trump hypothesis could be brought under debate. If there was indeed a strong social desirability bias against expressing support for Trump as the hypothesis claims, interviewer-administered polls (e.g., live phone polling

as seen by companies such as USC and TIPP) should record lower levels of Trump support than self-administered polls (e.g., online polling methodologies such as CVoter and SurveyMonkey) [8]. Post 2016 election analysis revealed estimates produced by live telephone polls were similar to those produced by self-administered online polling indicating to some extent that the shy Trump hypothesis may not have been as prevalent as initially thought.

6 Conclusion

Given a dataset containing results from 39 unique opinion polling companies, 7 variables spanning 264 separate surveys regarding the 2016 US election, exploration of the data has revealed that many polling companies differ in their practice and accuracy. Whilst practices span methodologies such as live phone calls, paper questionnaires and straw voting, the most prevalent form of opinion poll used in the 2016 US election was online surveys, with SurveyMonkey producing questionnaires for almost 400,000 voters - more than half the total sample frame seen from the 39 opinion polling companies in 2016. Whilst some companies indicate high levels of accuracy in their methodology output as seen prior by figure (5), pollsters outcomes deviate to a significant extent - IBD/TIPP polled a H:D output of 43%:42% with a margin of error of 3.1%. Companies output inaccurate and unstable results and thus their predictions must be interpreted with care.

Reasons for instability were analysed in the context of the 2016 US election yet without loss of generality, non-response bias and late deciding voters seem to be key factors producing uncertainties in polling outcomes. An example can be seen in work from Kennedy et al who states that there is "clear evidence that voter turnout changed from 2012 to 2016 in ways that favored Trump and other Republicans, though there is only mixed evidence that misspecified likely voter models were a major cause of the systematic polling error". For future polling procedures, rising issues discussed such as the "Fake News" epidemic produced by the advent of social media must be internalized within opinion polling company frameworks. Further research into this area such as investigating media sites likes Twitter and their algorithms influence on election outcomes could yield fruitful results in terms of understanding this upcoming polling instability issue.

References

[1] Peter Barnes. "reality check: Should we give up on election polling?". *BBC News*, 2016.

[2] Christina Bellantoni. How we did it: Usc dornsife/los angeles times poll methodology. *Los Angeles Times*, 2016.

[3] A. Bovet. Influence of fake news in twitter during the 2016 us presidential election. *Nature Communications*, 2019.

[4] Scott Clement. How the washington post-surveymonkey 50-state poll was conducted. *The Washington Post*, 2016.

[5] Philip E. Assessing the accuracy of polls and surveys. 1986.

[6] Ryan Finley. Surveymonkey,"board of directors,". 2016.

[7] Gallup. Understanding gallup's likely voter procedures for presidential elections. 2016.

[8] Courtney Kennedy. An evaluation of the 2016 election polls in the united states. *Public Opinion Quarterly, Vol. 82*, 2018.

[9] David Lauter. The usc/l.a. times poll saw what other surveys missed: A wave of trump support. *Los Angeles Times*, 2016.

[10] Editor Yashwant Deshmukh. Upi/cvoter pollstate tracker. 2016.