

Forecasting Congressional Elections Using Facebook Data

Matthew C. MacWilliams, *University of Massachusetts, Amherst*

ABSTRACT Facebook constantly tracks the growth of each congressional candidate's fan base and the number of people engaging with candidates online. These Facebook metrics comprise a rich dataset that theoretically may capture the effectiveness of campaigns in building participatory support as well as their potential to mobilize support. When added to electoral fundamentals similar to those used in national-election forecasting, can Facebook data be used to develop a reliable model for predicting vote-percentage outcomes of individual congressional contests? The results of an exploratory investigation reveal that fan participation and mobilization metrics tracked by Facebook produced surprisingly accurate election predictions in the 2012 US Senate races studied. The question remains, however, whether these results are a "flash in the 2012 pan" or an indication that using Facebook statistics to measure campaign effectiveness is a new tool that scholars can use to forecast the outcome of congressional campaigns.

Although election forecasting has provided an excellent proving ground for theories of voting behavior (Lewis-Beck and Stegmaier 2014; Linzer 2014; Sides 2014), some political scientists contend that forecasting is now at a crossroads. It is ironic that in the age of Big Data—in which many extol the power of web-generated data and imbue it with the ability to solve some of humanity's vexing problems (Lohr 2012)—it is a "lack of data" that limits the development of more accurate electoral-forecasting models (Linzer 2014).

Based on my experience with online media, I wondered whether the treasure trove of information produced online might provide a solution to the data problem that confronts forecasting.² In 2012, I hypothesized that Facebook metrics might be paired with electoral fundamentals in a simple model to predict the outcomes of individual congressional races and produce continually updated forecasts. Specifically, I asked: "Can candidate-page fan and engagement statistics tracked by Facebook be used to forecast congressional-campaign results?"

The initial answer to this question is: quite possibly. The performance of the Facebook Model in forecasting the vote in seven hotly contested campaigns for US Senate in 2012 indicates that readily available and transparent Facebook metrics—paired in a model with fundamental electoral benchmarks similar to those used in national-election forecasting—may

provide an accurate new tool for predicting the results of individual congressional contests.

The story of the Facebook Model's performance begins, as it should, with the proven election theories in which it is rooted and the one additional theoretical insight that makes it a promising hybrid.

THEORY

Lewis-Beck (2014) advances five fundamental theories of voting behavior that have been positively tested by forecasting. Four of the theories form the backbone of the Facebook Model. The model begins with the premises that voters are retrospective (Fiorina 1978; Lewis-Beck and Stegmaier 2000, 2014; Sides 2014); incumbency matters (Campbell 2014); and although "campaigns [can] influence...electoral outcome[s]," the partisan preferences of voters are not "easily swayed" (Lewis-Beck and Stegmaier 2014). The Facebook Model adds to the latter two premises the following simple theoretical insight:

Although voters' partisan preferences are not easily swayed, the willingness of partisans and fence sitters to publicly commit to and engage with a candidate can be influenced by the candidate's campaign. Furthermore, the more supporters that a campaign enlists and engages in political action before Election Day, the more likely that campaign is to win.

Two basic Facebook metrics can be used to quantify a campaign's success in enlisting and engaging Facebook users. Candidate "likes" track the number of Facebook users who enlist in a campaign by becoming fans of a candidate's page.³ Facebook's "people talking

Matthew C. MacWilliams is a teaching associate and PhD candidate in the department of political science at the University of Massachusetts, Amherst. He can be reached at macmatt@sprynet.com.

about this” (PTAT) statistic⁴ measures engagement by counting interactions between users and a candidate’s page. I argue that these Facebook measurements are a real-time measurement of a campaign’s effectiveness in enlisting and engaging supporters. When these data are included in a model with electoral fundamentals used in standard forecasting, the outcomes of individual Senate races can be predicted. In 2012, my team used this model to predict outcomes in seven US Senate races starting eight weeks before Election Day (MacWilliams and Erikson 2012).

FACEBOOK’S RELEVANCE TO FORECASTING

Why might Facebook data be a useful tool for estimating campaign effectiveness? Since the inception of social media use by candidates in 2006, research has found that political activity on Facebook mirrors offline political action.

First, in terms of enlistment, Facebook fans are described as “a proxy for the underlying enthusiasm and intensity of support a candidate generates” (Williams and Gulati 2007). A significant correlation between online fans and offline vote share was documented even when controlling for campaign expenditures, press coverage, and organizing (Williams and Gulati 2008).

Why might Facebook data be a useful tool for estimating campaign effectiveness? Since the inception of social media use by candidates in 2006, research has found that political activity on Facebook mirrors offline political action.

In terms of engagement, scholars have found a significant relationship between online and offline participation in which greater Facebook political activity is correlated with increased political action offline (Park, Kee, and Valenzuela 2009; Vesnic-Alujevic 2012) and is a “significant predictor of other forms of political participation” (Vitak et al. 2011). Political engagement on Facebook leads to “mobilizing political participation” offline (Feezell, Conroy, and Guerrero 2009).

The mobilizing effect of Facebook messages distributed peer-to-peer or en masse is also potent. A randomized test conducted in 2010 (N = 61 million) of third-party, get-out-the-vote Facebook messages found that they “directly influenced the voting behaviors of millions of Americans” (Bond et al. 2012).

Second, the Facebook data used in my model are standardized measurements that are readily accessible and regularly tracked. These data avoid many of the limits and methodological challenges found in many Big Data datasets, including Twitter (Boyd and Crawford 2011). The “right now” availability of Facebook data and resulting lack of historical record (Bollier and Firestone 2010), however, remain a challenge that can be surmounted only by capturing data weekly, as our team did during the closing weeks of the 2012 election and has continued to do since September 2013.

Third, Facebook is ubiquitous. In 2013, Pew Research reported that “Facebook is popular across a diverse mix of demographic groups” (Duggan and Smith 2014). Of those Americans who are online, 71% are on Facebook, 63% of whom check Facebook at least once a day. Moreover, 45% of Internet users 65 and older now use Facebook. This represents a 28-percentage-point growth in seniors’ use of Facebook in only one year (Project 2013). Facebook is no longer simply a social medium; it has become a social utility that campaigns are using to reach, activate, and mobilize

voters. Facebook users comprise neither a random nor a perfectly selected sample of the American electorate, and they are not conceptualized as such in the model. Instead, the relative effectiveness of campaigns in enlisting, engaging, and mobilizing Facebook users is theorized as a proxy for estimating the effectiveness of a campaign to generate support, activism, and votes among voters—much as Americans’ views of the economy in presidential forecasting models are used as a tool for estimating retrospective voting.

THE MODEL

Following the best practices of forecasting, the Facebook Model is steeped in theory, parsimony, and transparency. It is founded on the assumption that past election results and incumbency are fundamentals that play an important role in shaping electoral outcomes (Brody and Sigelman 1983; Campbell 2009; Campbell and Garand, 2000; Lewis-Beck and Rice 1992; Rosenstone 1983). The model adds to this foundation a participation variable (quantified through social media statistics generated by Facebook) that theoretically captures the effectiveness of each campaign’s efforts to enlist and engage voters, as well as their potential to mobilize voters on Election Day. The Facebook Model is specified as follows:

$$\text{Senate Vote} = f(\text{partisan voting index} + \text{incumbency} + \text{participation advantage})$$

In the model, Senate vote is the forecasted percentage of the two-party vote won by either major-party candidate. It is a function of the partisan vote index (PVI), which measures past election results, incumbency, and the estimated candidate-participation advantage generated from Facebook metrics.

PVI is estimated regularly by the *Cook Political Report* and has been used to undergird other election forecasts (Cook 2012). For example, Campbell’s 2012 House Seats-in-Trouble forecasting model used the *Cook Political Report’s* race-by-race analysis, which is predicated in large part on PVI (Campbell 2012b).

PVI averages the electoral performance of many candidates in a state or district over time to calculate existing partisan advantage. In this way, it captures the increasing polarization that presents statistical challenges to presidential models (Campbell 2014) but negates the fundamental advantages enjoyed by some Congressional incumbents.

The second fundamental variable—incumbency—is added to the Facebook Model to correct this PVI shortcoming. In presidential forecasting models, incumbency often is captured by a dichotomous variable. The inadequacy of quantifying presidential incumbency with a simple binary term is a contested question (Campbell 2014). Conceptualizing incumbency in a similar manner for individual Senate races—given the obvious electoral variations among Senate incumbents—is even more problematic. Thus, in the Facebook Model, incumbency advantage or disadvantage is determined by calculating how an incumbent performed, compared to the reported PVI, in the previous election. An incumbent Senator who won by five more percentage points in 2006 than predicted by the PVI would enjoy a five-percentage-point incumbent

advantage in the 2012 model—if the PVI had remained constant in the intervening years.⁵

The third variable, which enables the model to produce a forecast, trend data, and nowcast (Lewis-Beck and Stegmaier 2014), is participation. The participation-advantage variable is theorized as a real-time measurement of the effectiveness of each campaign in enlisting and engaging Facebook users as well as its potential to mobilize the vote on Election Day. These three measurements are designed to capture and quantify the Facebook effects identified and studied by scholars since the 2006 elections (Bond et al. 2012; Feezell et al. 2009; Vesnic-Alujevic 2012; Vitak et al. 2011; Williams and Gulati 2008; Williams and Gulati 2007, 2009a, 2009b; Zhang, Johnson, Seltzer, and Bichard 2010).

The first component of participation advantage is Facebook “likes.” Likes are a measure of Facebook users’ decisions about a candidate before Election Day. The growth of a candidate’s likes or fan base (i.e., Enlist Growth) over time tracks the effectiveness of a campaign in enlisting support among Facebook users.

The second component of participation advantage is Facebook’s PTAT statistic. PTAT measures active engagement with candidates, beyond mere support, in real time. Facebook users who engage with candidates online are politically mobilized.

The second component of participation advantage is Facebook’s PTAT statistic. PTAT measures active engagement with candidates, beyond mere support, in real time. Facebook users who engage with candidates online are politically mobilized. Although engagement with a candidate ebbs and flows depending on campaign events, success in building this politically mobilized group of activists over time (i.e., Engage Growth) is another component of campaign effectiveness.

The third component of participation advantage is the potential of a campaign to mobilize voters at a particular time (i.e., Mobilization Potential). This is conceptualized as the number of engaged PTATs divided by the campaign’s current fan base.

In the Facebook Model, these three measurements of campaign effectiveness are measured and combined weekly to produce the model’s dynamic participation-advantage variable (PA). How was this accomplished in 2012?

During each of the last nine weeks of the campaign, Facebook data for Senate candidates were collected and factored into the following equation to produce candidate participation scores (PS). For clarity, we use an example of a candidate named Hertz:

$$\text{Hertz Participation Score (PS)} = \text{Enlist Growth} \% * \text{Engage Growth} \% * \text{Mobilization Potential} \%$$

The participation score of Hertz’s opponent, Avis, also was calculated using the same formula.

Because Hertz and Avis are competing for votes from the same pool of voters, a Relative Participation Score (RPS) was calculated for each candidate, as follows:

$$\text{Hertz's RPS} = \text{Hertz PS} / (\text{Hertz PS} + \text{Avis PS}) * 0.1$$

$$\text{Avis's RPS} = \text{Avis PS} / (\text{Avis PS} + \text{Hertz PS}) * 0.1$$

From these two RPS figures, an absolute candidate PA was calculated each week by subtracting one candidate’s RPS from the other candidate’s RPS, as follows:

$$\text{Hertz's PA} = \text{Hertz's RPS} - \text{Avis's RPS}$$

Finally, to produce the weekly campaign forecast, the Hertz–Avis vote was divided equally first between the two candidates and then adjusted to account for the PVI,⁶ incumbency advantage, and weekly.

$$\text{Hertz's Senate Vote} = 50 + [\text{PVI}/2 + \text{Incumbency} + \text{PA}]$$

The vote percentage for Avis is simply 100 minus the Hertz estimated percentage.

2012 FACEBOOK MODEL PERFORMANCE

In 2012, candidate likes and PTAT data from September 1 to November 3 for major-party candidates in 15 of the most competitive Senate races⁷ were gathered daily using PageData.⁸ Competitive races were chosen for two reasons: (1) they provide a more difficult prediction challenge, and (2) competitive Senate campaigns

are more likely to use Facebook. The data from seven of the campaigns⁹ were complete during the entire period studied and were used to assess the performance of the model. PageData metrics for the other eight campaigns were incomplete and therefore excluded from our analysis.¹⁰ The model’s performance was assessed in the following three ways:

- accuracy of the weekly model forecasts immediately after Labor Day
- performance of the Facebook Model versus a model using only fundamental variables
- performance of the Facebook Model versus weekly aggregations of race-level polling data

Model Accuracy

To assess the accuracy of the Facebook Model predictions eight and seven weeks before Election Day, election results for the Senate races studied were converted to two-party candidate totals. These results provided the dependent variable against which the forecasted percentages produced by the model for the weeks ending September 14 and 21 were regressed.¹¹

The R-squareds for the first and second sets of Senate race predictions produced by the model for the weeks ending September 14 and September 21 were 0.772 and 0.746, respectively. Moreover, in both weeks, the Facebook Model accurately predicted the ultimate Senate victors.

Model Performance versus Fundamentals

The performance of the Facebook Model also was tested against a fundamentals-only model that used PVI and incumbency variables to produce predictions. In this test, if the Facebook Model

produced a higher R-squared than the fundamentals alternative, it added to the accuracy of the forecast.

In six of the eight assessment weeks, the Facebook Model (table 1) outperformed the fundamentals alternative. The two weeks in which the Facebook Model failed to outperform the alternative are an indication of the sensitivity of the model to relative changes in the performance of competing candidates. Averaging Facebook Model predictions over two weeks (i.e., a technique that is tested post-election that produces a rolling forecast) smoothes out the volatility, maintains the trending and nowcasting capability of the model, and produces forecasts that exceed the fundamental baseline every week.

Model Performance versus Poll-of-Polls Averages

Finally, the accuracy of Facebook Model predictions was evaluated against polling results—a benchmark suggested in the 2012 *PS: Political Science and Politics* Symposium (Campbell 2012a). First, the results of all 212 polls completed from September 8 to November 3, 2012, in the seven Senate campaigns under study were gathered from the *Huffington Post Election Dashboard* (HuffPost 2012) Starting with the week of September 8–14 and continuing through November 3, the results in each race were averaged, converted into two-party candidate totals to arrive at weekly poll-of-polls candidate estimates, and then regressed against election results.

Table 2 compares the weekly poll-of-polls R-squared to the Facebook Model. The simple Facebook Model was a better predictor of outcomes in the Senate races studied in five of eight weeks. It is important to note that the Facebook Model was a better predictor of election results in four of the five weeks that were farthest from Election Day. In other words, when compared to poll-of-poll averages, the Facebook Model was better at forecasting outcomes the farther the prediction was from Election Day.

Table 2

R-Squared of the Facebook Forecasting Model Predictions Versus R-Squared of Averaged Polls-of-Polls Predictions

	Facebook Forecast Prediction	Averaged Poll-of-Polls Prediction
	R-Squared	R-Squared
Week Ending Sept 14	0.7718	0.2182
Week Ending September 21	0.7458	0.7452
Week Ending September 28	0.5356	0.8190
Week Ending October 5	0.7893	0.0965
Week Ending October 12	0.5513	0.4514
Week Ending October 19	0.7161	0.5303
Week Ending October 26 <i>Standard Error</i>	0.7677	0.8393
Week Ending Nov 3 <i>Standard Error</i>	0.7823	0.8810
	N=7	N=212

The dependent variable is two-party vote.

CONCLUSION

Election forecasting is a worthy pursuit that has tested the mettle of many voting-behavior theories (Sides 2014). Yet, as discussed in the 2014 *PS: Political Science and Politics* Symposium articles, it experiences several challenges, including lack of data, lack of timeliness, distance from the campaign narrative, inadequate specification of incumbency, partisan polarization, and national-level aggregation of results (Campbell 2014; Lewis-Beck and Stegmaier 2014; Linzer 2014; Sides 2014).

Since its inception, election forecasting has focused on the grand prize: predicting the outcome of presidential elections. However, in 2012, only 7 of the 12 regression models highlighted in *PS: Political Science and Politics* predicted the reelection of President Obama (Campbell 2012b; Lewis-Beck and Stegmaier 2014).

For the study of election forecasting to progress, more cases for experimentation and more reliable data sources are needed. The Facebook Model is an attempt to answer both of those needs.

The results of this exploratory investigation indicate that Facebook likes and PTAT metrics, when added to standard forecasting fundamentals, can produce surprisingly accurate vote forecasts in individual contests. The question remains, however, whether these results are an anomaly or a tool to expand the statistical forecasting of election results to campaigns for Congress. Only time and the testing of the model in future elections will determine if Facebook metrics are indeed a new tool to add to the forecasting toolbox. ■

NOTES

1. Big Data, however, is not a silver-bullet solution for the lack of data confronted by forecasters. The pitfalls and limits of using Big Data are well documented (Bollier and Firestone 2010; Boyd and Crawford 2011).
2. My colleague, Edward Erikson of the University of Massachusetts, Amherst, participated in the initial development of this model. He coauthored articles in the popular press on this subject with Nicole Luna Berns and me. Berns is my research assistant, courtesy of the University of Massachusetts, Amherst.

Table 1

R-Squared of the Facebook Forecasting Model Predictions Versus R-Squared of Predictions Based on Static Fundamentals (PVI and Incumbency)

	Facebook Forecast Prediction	Static Fundamental Prediction
	R-Squared	R-Squared
Week Ending September 14	0.7718	0.5831
Week Ending September 21	0.7458	0.5831
Week Ending September 28	0.5356	0.5831
Week Ending October 5	0.7893	0.5831
Week Ending October 12	0.5513	0.5831
Week Ending October 19	0.7161	0.5831
Week Ending October 26	0.7677	0.5831
Week Ending November 3	0.7823	0.5831
	N=7	

The dependent variable is two-party vote.

3. Page “likes” is a long-standing Facebook measurement that predates the launching of the first candidate pages in 2006. “Likes” are the number of unique Facebook users who click the “like” button on a Facebook page. Those who “like” a page are said to be “fans” of the page.
4. Facebook first began reporting PTAT statistics in October 2011. PTAT is updated daily and averaged over a rolling seven-day period. Multiple interactions by a user during one seven-day period are counted only once. Counted interactions include liking a page, liking a post, commenting on a post, sharing a post, posting on the page’s wall, answering a question asked on the page, RSVPing to an event, tagging the page in a photograph, and mentioning the page in a post (SocialTimes, 2012).
5. Any changes in the PVI between 2006 and 2012 also were added to incumbency to stay current with the polarizing trends that Campbell contends are a challenge to fundamentally based forecasting (Campbell 2014).
6. The PVI is divided by 2 to account for the 50–50 split of the vote among the two-party candidates.
7. As rated by the *Cook Political Report* (Cook 2012).
8. PageData is a company that tracks “user interaction and engagement data for millions of Facebook Pages.”
9. The seven competitive Senate races with complete data included campaigns in Connecticut, Florida, Indiana, Massachusetts, Nevada, Ohio, and Virginia.
10. PTAT data comprised the main missing-data culprit in 2012—probably because the measurement was first introduced by Facebook in late 2011.
11. The week ending September 7 was the baseline for the Facebook measurements.

REFERENCES

- Bollier, D., and C. M. Firestone. 2010. *The Promise and Peril of Big Data*. Washington, DC: The Aspen Institute.
- Bond, R. M., C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler. 2012. “A 61-million-person experiment in social influence and political mobilization.” *Nature* 489: 295–298.
- Boyd, D., and K. Crawford. 2011. “Six Provocations for Big Data.” *Social Science Research Network*. Available: http://ssrn.com/abstract_1926431.
- Brody, R., and L. Sigelman. 1983. “Presidential Popularity and Presidential Elections: An Update and Extension.” *Public Opinion Quarterly* 47 (3): 325–328.
- Campbell, J. E. 2009. “An Exceptional Election: Performance, Values, and Crisis in the 2008 Presidential Election.” *Presidential Studies Quarterly* 40 (2): 225–246.
- . 2012a. “Forecasting the 2012 American National Elections.” *PS: Political Science and Politics* 45 (4): 610–613.
- . 2012b. “Forecasting the Presidential and Congressional Elections of 2012: The Trial-Heat and the Seats-in-Trouble Models.” *PS: Political Science and Politics* 45 (4): 630.
- . 2014. “Issues in Presidential Election Forecasting: Election Margins, Incumbency, and Model Credibility.” *PS: Political Science and Politics* 47 (2): 301–303.
- Campbell, J. E., and J. C. Garand. 2000. “Forecasting US National Elections.” *In Before the Vote: Forecasting American National Elections*, 3–16. Thousand Oaks, CA: Sage Publications, Inc.
- Cook, C. 2012. *The Cook Political Report*.
- Duggan, M., and A. Smith. 2014. “Social Media Update 2013.” Pew Research Center, December 30, 2013. Available online: <http://www.pewinternet.org/2013/12/30/social-media-update-2013>.
- Feezell, J., M. Conroy, and M. Guerrero. 2009. “Facebook is... Fostering Political Engagement: A Study of Online Social Networking Groups and Offline Participation.” Paper Presented at 2009 APSA Annual Meeting.
- Fiorina, M. P. 1978. “Economic Retrospective Voting in American National Elections: A Micro-Analysis.” *American Journal of Political Science* 22 (2): 426–443.
- HuffPost. 2012. *HuffPost Pollster Senate Elections 2012*.
- Lewis-Beck, M. S., and T. W. Rice. 1992. *Forecasting Elections*. Washington, DC: CQ Press.
- Lewis-Beck, M. S., and M. Stegmaier. 2000. “Economic Determinants of Electoral Outcomes.” *Annual Review of Political Science* 3 (1): 183–219.
- . 2014. “US Presidential Election Forecasting.” *PS: Political Science and Politics* 47 (2): 284–288.
- Linzer, D. A. 2014. “The Future of Election Forecasting: More Data, Better Technology.” *PS: Political Science and Politics* 47 (2): 326–328.
- Lohr, S. 2012. “The Age of Big Data.” *New York Times*, February 11, 2012. <http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html>.
- MacWilliams, M., and E. Erikson. 2012. “The Facebook Bump.” *Politico*, November 18, 2012. <http://www.politico.com/news/stories/1112/83572.html>.
- Park, N., K. F. Kee, and S. Valenzuela. 2009. “Being Immersed in Social Networking Environment: Facebook Groups, Uses and Gratifications, and Social Outcomes.” *CyberPsychology & Behavior* 12 (6): 729–733.
- Project, P. I. R. (2013). *Social Media Update*.
- Rosenstone, S. J. 1983. *Forecasting Presidential Elections*. New Haven, CT: Yale University Press.
- Sides, J. 2014. “Four Suggestions for Making Election Forecasts Better, and Better Known.” *PS: Political Science and Politics* 47 (2): 339–341.
- SocialTimes. (2012). Retrieved from www.insidefacebook.com
- Vesnic-Alujevic, L. 2012. “Political participation and web 2.0 in Europe: A case study of Facebook.” *Public Relations Review* 38 (3): 466–470.
- Vitak, J., P. Zube, A. Smock, C. T. Carr, N. Ellison, and C. Lampe. 2011. “It’s complicated: Facebook users’ political participation in the 2008 election.” *CyberPsychology, Behavior, and Social Networking* 14 (3): 107–114.
- Williams, C., and G. Gulati. 2008. “What Is a Social Network Worth? Facebook and Vote Share in the 2008 Presidential Primaries.” Paper presented at 2008 APSA Annual Meeting.
- Williams, C. B., and G. J. Gulati. 2007. “Social Networks in Political Campaigns: Facebook and the 2006 Midterm Elections.” Paper presented at the 2007 annual meeting of the American Political Science Association.
- . 2009a. “Explaining Facebook Support in the 2008 Congressional Election Cycle.” *Working Paper* No. 26.
- . 2009b. “Facebook Grows Up: An Empirical Assessment of its Role in the 2008 Congressional Elections.” Paper presented at the 2009 Annual Meeting: Midwest Political Science Association.
- Zhang, W., T. J. Johnson, T. Seltzer, and S. L. Bichard. 2010. “The Revolution Will be Networked: The Influence of Social Networking Sites on Political Attitudes and Behavior.” *Social Science Computer Review* 28 (1): 75–92.