# A Bottom-Up Approach to Linguistic Persuasion in Advertising:

## Predicting and explaining the effects of TV ads using automated text analysis

Nick Beauchamp[*]
Columbia University

September 18, 2012

**Abstract**

This paper presents a new, bottom-up approach to modeling the effects of television advertising on vote intention. Because ads are so numerous and varied, existing studies generally begin with a specific theory of persuasion, or must simplify the data down to a few latent dimensions or effective ads. Instead, this new approach first develops a one-at-a-time regression technique to estimate the effects of hundreds of different ads on vote intention during the 2004 presidential campaign. The aggregate effect of advertising is found to be significant, though many individual ads have small or backfiring effects. To explain these varying effects, new automated text analysis procedures are developed which can predict the effects of ads based only on their text, and reveal complex and asymmetric strategies that mix affect, policies, issue ownership, negativity, and targeting. This bottom-up procedure constitutes a new method for understanding persuasion in campaigns and politics more broadly.

[*] Email: nb2389@columbia.edu; web: nickbeauchamp.com. For their helpful comments on this paper or a presentation of it, the author would like to thank: Jonathan Nagler, Michael Laver, Nathaniel Beck, Arthur Spirling, Gary King, Burt Monroe, Adam Bonica, Alex Herzog, the participants at the Saint Louis Area Methods Meeting, and the faculty and students at NYU.

# 1 Introduction

How does political advertising affect vote intention? The last few decades have seen a proliferation of explanations, ranging from general theories of persuasion to theories specific to US presidential campaigns in the 21st century. On the broadest level, theories of persuasion range from behavioral and emotional at one end[1] to rational models based on utility and information updating at the other,[2] while in between these poles might lie the interrelated theories of agenda-setting, framing, and priming.[3] On a level more specific to campaigns, we find large literatures focusing on the effects of negative advertising,[4] or examining the strategic differences between incumbents and challengers,[5] or probing the tradeoffs between "issue ownership" and issue convergence or "riding the wave" of topical issues.[6] And at the most granular level are topics specific to a given campaign in a given year: in 2004, these might include the Iraq war, John Kerry's service in Vietnam, health care, taxes, the deficit, etc. Deepening the complexity, explanations at different levels may overlap

[1] For a few prominent examples from this rapidly growing literature, see Mutz (1996), Brader (2006), Mutz (2007), Westen (2007) and Ridout & Franz (2011).

[2] For important examples from this larger and somewhat earlier literature, see Lodge, McGraw & Stroh (1989), Ferejohn & Kuklinski (1990), Popkin (1991), Franklin (1991), Page & Shapiro (1992), Zaller (1992), Sniderman, Brody & Tetlock (1993), Gelman & King (1993), Bartels (1996), Alvarez (1998), Holbrook (1999) and Lupia & McCubbins (2000).

[3] For agenda setting, see for instance McCombs & Shaw (1993), McCombs, Shaw & Weaver (1997) and Jacobs & Shapiro (2000). For priming, see for instance Iyengar & Kinder (1987) and Krosnick & Kinder (1990). For framing, see for instance Pan & Kosicki (1993), Iyengar (1994), Druckman (2001*b*), Druckman (2001*a*), and Druckman (2004).

[4] See, for just a few examples, Ansolabehere & Iyengar (1995), Kahn & Kenney (1999), Stevens (2005), Geer (2006), Lau, Sigelman & Rovner (2007), Franz & Ridout (2007), Stevens, Sullivan, Allen & Alger (2008), Jackson, Mondak & Huckfeldt (2009), and Krupnikov (2011).

[5] Research in this vein has naturally emphasized congressional and local elections (Jacobson 1978, Mann & Wolfinger 1980, Ansolabehere, Snowberg & Snyder 2006, Benoit & Marsh 2008), particularly regarding spending (Abramowitz 1991, Ansolabehere & Gerber 1994, Krasno & Green 1988, Moon 2006), but there are also numerous examinations of incumbency effects on the efficacy of presidential campaigns (Bartels 1993, Alvarez 1998, Holbrook 1999, Prior 2007, Ridout & Franz 2011).

[6] These two strategies are often discussed in conjunction with each other, and obviously relate to agenda setting and priming as well. Issue ownership research begins with Riker and proceeds robustly from there (Riker & Ordeshook, P.C. 1968, Petrocik 1996, Iyengar & Valentino 2000, Simon 2002), with many examining the tension between issue convergence and divergence (Ansolabehere & Iyengar 1994, Spiliotes & Vavreck 2002, Brasher 2003, Sulkin & Evans 2006, Vavreck 2009), although some recent research suggests that ownership effects, or even attempts by campaigns, may be quite weak (Sigelman & Buell Jr 2004, Damore 2004, Damore 2005, Sides 2006, Kaplan, Park & Ridout 2006, Sides 2007, Sides & Karch 2008).

and interact, so that, for instance, negative ads might operate via information as well as emotion,[7] or might affect challengers more than incumbents, or might have been unusually effective in the infamous 2004 Swift Boat ads. And all of this is assuming that, of the hundreds of intended and inadvertent persuasive strategies, any of them at all have discernible effects – far from the scholarly consensus.[8]

How are we to deal with this welter of simultaneous and interacting mechanisms? One approach is to focus on testing a specific theory of persuasion, at whatever level of generality, and ignore all the others. But although an experimentally-crafted advertisement might ideally vary along a single persuasive modality, in the real world every ad encompasses numerous different interacting strategies and ideas. You may think you are measuring differences in negative affect, but you could also be measuring differences in information; or you think you are measuring issue convergence, when it is really challenger focus or previous agenda-setting. And that is just for a single ad; in reality, all these themes are interacting with all of those in every other ad being broadcast. It may still be the case that a few identifiable strategies do most of the persuasive work across all ads in a given campaign. But if that is not the case – if the themes and strategies that affect viewers are as numerous and diverse as the ads are themselves, or more so – then other methods are necessary to model the persuasive effects of a campaign.

Taking television advertising in the 2004 presidential campaign as an excellent case study, this paper shows that persuasion in advertising is best modeled by an approach that tackles the full complexity of the advertising environment. The traditional theory-first, top-down approach,

[7] See, for instance, Stevens (2005) Geer (2006), and Franz & Ridout (2007).

[8] As Brady, Johnston & Sides (2006) put it, "The prevailing scholarly consensus on campaigns is that they have minimal effects. Minimal effects mean in essence minimal persuasion." They note that this holds especially for presidential campaigns, and has been the view at least since Lazarsfeld, Berelson & Gaudet (1948), although more recent work has found small but consistent effects; see Holbrook (1996) or Campbell (2008), along with many of the other works here.

where one categorizes ads according to whether they are negative or health-related (for instance) and models the effect of total negative or health-related broadcasts on vote intention, works far less well than an approach that models the full complexity of advertising. The bottom-up approach, by contrast, begins with hundreds of variables, each counting the number of broadcasts of each unique ad; models the effect of all of those ads; and then uses automated methods to move upward from those inferred effects plus the textual content of the ads, to the myriad themes and strategies all at work at once in the campaign. The difficulty of this approach lies in modeling a system with hundreds of independent variables, whether those variables are the ads themselves (as discussed in Stage One of this paper), or the words and themes within those ads (as discussed in Stage Two). This is not merely a methodological challenge, but more importantly a theoretical problem about the structure of causality in complex systems. If the top-down approach fails, is it merely a matter of identifying a better set of underlying strategies that explain the variation among ads? And if not, if there is no detectable simplification, is it instead the case that the effects of ads are so idiosyncratic that we should instead just identify a few that happen to do most of the work (such as the infamous Swift Boat ads)? By comparing methods suited to each of these causal structures, we will see that, while some seem to model persuasion in advertising better than others, best of all is a new approach, called here one-at-a-time regression, that models the effect of every ad separately without simplifying or reducing the system of hundreds of ads, tiny though many of their effects may be.[9]

Once this model of persuasion has been shown to be superior to the alternatives via out-

[9] In many ways, this problem is analogous to measuring the genes associated with a complex disease: earlier candidate-driven approaches consisted of hypothesizing and testing each gene that was suspected of helping to cause a disease, but this is hugely inefficient. Better than laboriously checking each hypothesis is an approach that measures all the gene activities that differ between the diseased and healthy individuals, extracts the full array of contributors all at once, and only then do we try to understand the complex mechanisms implied by this array. See Haibe-Kains, Desmedt, Sotiriou & Bontempi (2008), Annest, Bumgarner, Raftery & Yeung (2009) or Bøvelstad, Nygård & Borgan (2009) for work in bioinformatics that closely resembles the Stage One procedures here.

of-sample testing, we still must explain what makes some of the ads more persuasive than others. Stage Two confronts the problem of complexity at another level: instead of hundreds of ads, we now have thousands of words, concepts, and themes at play in each ad, all of which may be contributing to the effects of those ads. For this stage, a number of new techniques in text analysis are developed or adapted, each again suited to a different potential causal structure. Perhaps every word and the ideas it evokes affects viewers independently, as the ads did; or perhaps content affects viewers in highly, but not entirely, idiosyncratic ways, so that the effect of an ad could be predicted by averaging the effects of textually similar ads with known effects. As we will see, a model that allows for both of these structures works better than either individually, showing that though complex, the effects of content on persuasion can still be modeled from the bottom up. And once this has been shown (as before, through out-of-sample testing), we can then use exactly these methods to discern which words and concepts are associated with the most effective ads.

The results of this two-stage procedure yield insights not just into general questions about the causal structure of complex persuasive systems like political advertising, but also into substantively important questions about the 2004 campaign and US presidential campaigns more generally. The 2004 campaign is particularly interesting because it was a tightly contested election which has been thoroughly examined with many suggestions but little consensus about what did or did not affect vote share, if anything did.[10] As we will see, Stage One reveals that the overall effect of advertising appears to have been small though significant (on the order of a few points in the pro-Kerry direction), but this small overall effect is due less to the ineffectualness of advertising or

[10] For instance, Kaid (2005) finds little effect of TV advertising on vote intention, though some reduction in political cynicism; Shaw (2006) shows that the effects from both sides of the campaign are large in the short term but largely cancel out in the longer term; Ridout & Franz (2011) show that only negative Republican ads (versus negative Democratic or positive ads) and fear- or anger-based Democratic ads (versus fear- or anger-based Republican ads or enthusiasm-based ads) were effective in 2004; and Hillygus & Shields (2008) find that micro-targeting was particularly effective in 2004 on the Republican side.

because the two sides cancelled each other out, than to poor ad deployment and emphasis by the campaigns; had one side better chosen which ads to emphasize, they could have greatly improved their outcome. But Stage Two is where the more detailed insights into this campaign are developed, revealing the diverse and highly asymmetric strategies underlying not what the two parties *intended* to do, but what they actually succeeded in doing. The results are complex and not amenable to any one of the dominant theories. The Democratic side was most successful when blanketing many areas with high-frequency, low-effect ads all making policy-based arguments emphasizing health care and prescription drugs. By contrast, the Republican side was more successful with a larger variety of relatively low-frequency, targeted ads making more negative and emotional attacks on Kerry on topics such as foreign policy, abortion, Kerry's Senate voting record, and of course his time in Vietnam. In accord with theories emphasizing the more fluid reputation of the challenger, both sides seemed more focused on Kerry, though his successful ads were more prospective, while Bush's attacks were more retrospective. And regarding issue ownership versus convergence, we will see that ownership seemed more dominant, where strategies that ventured into the other's territory were more likely to backfire with opposite their intended effects. While it does take expertise and knowledge of the campaign context to understand the implications of the word scaling, the complexity and asymmetry of the successful strategies could never have been discerned using even the most well-crafted top-down procedure. Fundamentally, this bottom-up approach offers a new procedure for modeling, predicting, and understanding effects in complex systems with hundreds or thousands of interacting treatments, as is common wherever words, meanings, and persuasion occur.

## 2 Data and groundwork

Throughout the 2004 campaign, the National Annenberg Election Survey (NAES) fielded a rolling national survey, and the Wisconsin Advertising Project (WiscAds) in conjunction with the Campaign Media Analysis Group (CMAG) recorded the time and location each television advertisement that was broadcast, along with "storyboards" with transcripts of each ad. In addition, WiscAds/CMAG has categorized each ad according to many of the theories above (eg, whether it is negative, is about women, is policy-oriented, is about the economy, etc.), and indeed a minor industry has developed using these categorizations to test various theories.[11] For the present purposes, the bulk of the 2004 presidential campaign is taken to cover the 16 weeks leading up to the election. The NAES provides individual level survey data on vote intention and various demographic characteristics used as controls;[12] the dependent variable is vote intention, and the independent variables of interest are the 359 ad count variables, plus the controls.

This constitutes a time-series multilevel model, with about 40,000 survey observations collected over about 100 days, and broadcast data that record the number of broadcasts of each unique ad in a region (Nielsen Designated Market Areas, DMA) that day. Although that model was tested, in the following analysis the data are aggregated by week and region, for a number of reasons. First, because the model must be estimated hundreds of thousands of times, and the multilevel version takes four or five orders of magnitude longer to run even on a computer cluster. Second, the essential independent variable, the ad broadcasts, are on the regional level and tend

[11] For just a few recent examples, see Stevens (2005), Sides (2006), Ansolabehere, Snowberg & Snyder (2006), Franz & Ridout (2007), Sides & Karch (2008), Krasno & Green (2008), Ridout & Franz (2011). See also http://wiscadproject.wisc.edu/publications.php for a more comprehensive list. Of course, simply testing all these theory-based categories, whether jointly or separately, runs into severe problems with spurious results, and one often encounters proclamations such as "In writing this book, we estimated hundreds of statistical models" (Ridout & Franz 2011). With so many possible theories to test, selection bias becomes a serious problem, hence the need for the present approach.

[12] Specifically, party ID, ideology, age, employment status, income, education, race, religiosity, south, and urbanism.

to run for a week or so, so the region-week is the natural level of analysis for advertising effects.[13] And while individual-level variations within regions may provide slightly better estimates of effect sizes, in this case the ad-effect estimates are very similar whether estimating the aggregate or multi-level models.[14] It will of course be interesting in the future to look at interactions between individual-level qualities and ad effects,[15] though this may be beyond the reach of the data at hand.

Thus for each observation, the dependent variable becomes intended Kerry vote as a proportion of the intended two-party vote in that week-region, while each control is the mean survey response for that week-region (eg, mean age or unemployment level in that region, proportion of blacks in that region, etc.). And each of the 359 ad count variables measures the total number of broadcasts of that unique ads in a region over a week.[16] At this level of aggregation, the data are still time-series cross-sectional, but each observation represents only about 50 survey responses. As a consequence, estimated effects for the independent variables are the same whether one uses at time-series cross-sectional estimator, or if one treats the data as pooled, with 859 observations.

[13] Fixed effects for regions are not included (as dummies or in a TSCS model). This is for a number of reasons: First, with 50+ regions and only 16 time units, the fixed effects would absorb most of the variance we might try to explain. But more importantly, including the fixed effects worsens out-of-sample prediction of the DV (using the procedures described below), showing that fundamentally, they do not improve model fit, and are better omitted.

[14] As Cameron & Trivedi (2005) put it in their *Microeconomics*, "in many applications with aggregate proportions data, such as unemployment rate by region, there is no desire to estimate individual-level parameters...Then the linear regression may be fine." (482). Using a log-proportion dependent variable or the heterogeneity corrections they suggest makes no significant difference in estimated effects, mainly because the data are noisy and the proportions generally near 0.5. And correcting for clustered standard errors here in unnecessary, since the error measurements are not used, in favor of out-of-sample testing. For further aggregation, see also Krasno & Green (2008), with critique by Franz, Freedman, Goldstein & Ridout (2008).

[15] Many of the studies discussed here examine interactions between persuasive effects and individual-level qualities like partisanship or knowledge. Prior (2007) and Hillygus & Shields (2008) provide particularly sophisticated examples of this, but though interesting, these mechanisms are beyond the scope of the present study.

[16] There have been extensive debates recently about how best to measure ad exposure. Although it might appear that one might better measure ad exposure with more than mere broadcast counts (Prior 2007, Huber & Arceneaux 2007), weighting by cost, TV program popularity, or more generally Nielsen Gross Rating Points (GRP), there are serious drawbacks with that approach. GRP measures, after all, are survey-based, and thus introduce additional measurement error and serious endogeneity problems, as pointed out by Franz & Ridout (2007), Vavreck (2007), Franz et al. (2008), Stevens et al. (2008) and Krasno & Green (2008) among others. Ridout, Shah, Goldstein & Franz (2004) and Krasno & Green (2008) find that weighted and unweighted measures correlate very highly; the latter find very similar results either way, though the former find small differences. When Vavreck (2007) examines similar differences in a more controlled setting, she finds that recall-based measures tend to produce less dependable results.

Tests of serial correlation show that at this level of aggregation, there is no autocorrelation between region observations from one period to the next, nor is there autocorrelation between residuals from a pooled model regressing vote intention on controls, nor do the estimates from the pooled model differ from those using standard time-series models.[17] Thus, at least at this level of aggregation, the data can be treated as a single pool,[18] a final simplification which also makes the hundreds of thousands of estimations necessary in the following models computationally feasible.[19]

## 3 Stage One: Estimating ad effects with one-at-a-time regression

Which causal structure best describes how ads affect viewers? Is it via a few idiosyncratic ads, a latent dimension shared by all ads, a few familiar themes or strategies, or via a myriad of smaller effects present in small and varying degrees in almost every ad? This is a deep question about how persuasion works, and the best way to evaluate this question is to compare the performance of methods suited to each causal structure. The categorization approach is the simplest: just include the CMAG categories as independent variables,[20] though since there are dozens of these, decisions need to be made about which to include. Reducing the set of 359 ad count variables to a few underlying dimensions is nearly as easy using established methods like principal component

[17] ARIMA models were explored for a number of different autoregressive, integrated, and moving average values (Moffitt 1993).

[18] Johnston, Hagen & Jamieson (2004) find that media effects generally last only for the week the ad is run in. Gerber, Gimpel, Green & Shaw (2011) "find relatively little evidence of time-series dynamics. Simple models, in other words, lead to roughly the same substantive conclusions as more elaborate models... Television's effects appear to peak during the week in which the advertisements air." Shaw (2006) pools time-series data for similar reasons, and Stevens et al. (2008) stresses the steeply declining effects of a given advertisement over time. Had there been temporal correlations, the data could also have been detrended in various ways to better allow pooled testing, but this was not necessary.

[19] Again, because of the modularity of this approach, in addition to additional controls or instruments, rapidly increasing computer power should make out-of-sample testing with the full multilevel model feasible in only a few years.

[20] Which, recall, count the number of times all ads of a certain category, such as negative, policy-rebuttal, or women-related, are broadcast in a given region during a given period. Because the parties presumably desire opposite effects, each of these category variables is split into two, according to who sponsored the ads.

analysis (PCA). And if one seeks a small subset of ads which are responsible for almost all of the persuasive effects, there are a number of algorithmic approaches that are more systematic than stepwise addition or subtraction; a popular and powerful new method is LASSO (least absolute shrinkage and selection operator), which shrinks most of the estimated coefficients to zero, leaving a small subset of purportedly significant variables (Tibshirani 1996).[21]

If on the other hand, one suspects that every ad might have effects, however small, which are nevertheless not summarizable with a few latent dimensions, then standard approaches are less useful. Because there are 359 unique ads and only 859 week-region observations, estimating individual ad effects via a traditional multiple regression would lead to extreme over-fitting. That is, we might get a high $\mathrm{R}^2$ and many apparently significant effect coefficients on our ad variables, but were we to use such coefficients to predict vote intention out-of-sample, the prediction would be very poor (as we will see). Instead, I develop a new method called one-at-a-time regression to tackle this problem. It allows us estimate the effect of every ad, a characteristic which will also be essential for text-based inferences later on. But equally important, it simply works better than the alternatives, showing that this model of persuasion fits the data better than the other causal models.

The one-at-a-time procedure is straightforward: rather than estimate $y = \mathbf{X}\boldsymbol{\beta}$ on all K independent variables at once, one instead makes K univariate OLS estimates:[22]

[21] This method is also similar to ridge regression and least angle regression; in practice, all of these approaches produce similar results, shrinking the least significant variables out of a large set. LASSO begins with the usual OLS coefficients, but seeks to minimize the following:

$$\arg\min_{\boldsymbol{\beta}} \|Y - X\boldsymbol{\beta}\| - \lambda \sum_{k=1}^{j} |\beta_k|$$

That is, the total size of the beta coefficients is constrained, which leads to many of the insignificant ones being shrunk to zero.

[22] In this regard it resembles the first step in many genomic approaches. For instance, in supervised principle component analysis (Friedman, Hastie & Tibshirani 2009) one begins with univariate estimates but then selects only the most significant to work with, whereas here using all of them works best.

$$y = \beta_k X_k + \mathbf{Z}\mathbf{\Gamma}_k \tag{1}$$

where $X_k$ is an ad count variable equal to the total number of broadcasts of that unique ad in that region during that week, and $\mathbf{Z}$ are any control variables. The procedure is inspired by the recent success of ensemble methods (Dietterich 1997, Dietterich 2000, Schapire 2003, Sebastiani 2002, Page 2008), where numerous studies have now demonstrated the ability of ensembles of models to out-perform single, more complex models, often because the ensembles can capture diverse characteristics of the data without over-fitting.[23]

Before testing these differing approaches on the ad data, there are a number of methodological concerns about the one-at-a-time approach that should be confronted. First, a clear concern with this approach is omitted variable bias, since from each regression almost every other variable is omitted! This is discussed in Appendix A, which shows that, although we can analytically correct for omitted variable bias when we know what variables are omitted, the correction actually worsens performance in many cases, since it is precisely by ignoring the concurrent effects of the other variables that the one-at-a-time approach can out-perform basic regression in these noisy, highly multi-variate circumstances. The second question is in what theoretical circumstances we would expect one-at-a-time to out-perform other models suited to highly multivariate data. This is examined Appendix B, where Monte Carlo simulations show that one-at-a-time works better than multiple regression and LASSO in very specific circumstances: when each independent variable has a small contribution to the dependent variable, and the signal-to-noise ratio (eg, the R$^2$) is low –

[23] Part of the advantage of these methods is that they ignore the sorts of correlations between independent variables that methods like regression take into account, often to their detriment when the data provide insufficient information to estimate these second-order effects. However, even the ensemble approach can over-extend itself: for instance, while ensembles of sub-models within a single model (eg, random forests) sometimes improve on a single model, these tools also run into problems with over-fitting with real-world data (Breiman 2001, Segal 2004).

just the circumstances we expect may obtain with political advertising. Finally, there is the more general issue of causal identification and endogeneity. This is not the place to adjudicate between experimental and observational approaches,[24] but examining variations in hundreds of ads across many small regions avoids many of the causal identification problems associated with observational studies (Krasno & Green 1988, Erikson & Palfrey 1998) which examine only coarse measures such as overall spending, state-level effects (Shaw 1999, Shaw 2006), or broad strategies like "negativity." Telling reverse-causation stories for variations in content is still quite possible – and indeed, the campaigns themselves tell them all the time – but as we will see, the connections between ads and effects are so erratic and tenuous, individually, that it is very difficult to argue that they reflect sophisticated modulations of campaign messages in response to slight changes in a candidate's local performance, rather than a campaign that may think it is doing a lot, but is mainly just throwing a lot against the wall, and only rarely correctly identifying what sticks.[25] But most importantly, this is a modular approach: instruments, natural experiments, or even field-experimental data could in the future be used for the Stage One estimation in place of direct observational data, although gathering such data in sufficient breadth and variety will be a challenge.

[24] For a few excellent laboratory studies of persuasion, see Iyengar, Peters & Kinder (1982), Ansolabehere & Iyengar (1995), Mutz (1998), Mendelberg (1997), Iyengar & Valentino (2000), Brader (2006). For critiques of their external validity and generalizability, see Brady, Johnston & Sides (2006) and Davenport, Gerber & Green (2010). Field experiments may be ideal, but lacking a hundred-million-dollar budget to properly randomize hundreds of different ads blanketing dozens of states over many months (Gerber et al. 2011), observational data remains quite useful for examining complex real-world persuasion, and as Brady, Johnston & Sides (2006) point out, even a successful field experiment generally only "identifies a potential rather than an actual effect." Natural experiments, while an excellent and less expensive alternative (Ansolabehere, Snowberg & Snyder 2006, Huber & Arceneaux 2007), can be at least as rare as a million-dollar research budget, and occasionally as troubled by unobserved covariates as pure observational studies. Notably, field and natural experiments as a rule tend to find much weaker effects than does laboratory work. Finally, even recent methodological improvements to observational data, such as matching (Imai 2005), have met with strong critiques from the experimental side (Arceneaux, Gerber & Green 2006).

[25] There are also various game-theoretic reasons for doubting the "master-strategist" campaign narrative. Even when it comes to raw spending levels, the campaign advertising competition is a Colonel Blotto resource-allocation game, a game which famously has no perfect equilibria, and for which complex mixed equilibria have only recently been characterized (Robson 2005) – and only then for the single-shot game, not this extensive one taking place over many weeks. Given the complexity of the game (Erikson & Palfrey 2000), the idea that correlations between hundreds of different ads and vote-intention derive from the effect of the former on the latter is considerably more plausible than the reverse.

### 3.1 Testing the models

When the models are as different as these, there is really only one uniform way to test them all on the same playing field, out-of-sample prediction: fit each model on a subset of the data, and predict the dependent variable for the rest.[26] For most of these approaches this is straightforward enough: regress vote intention on the category variables, PCA components, or LASSO results (plus controls), and then predict in the usual OLS fashion. But for one-at-a-time regression, we need a method for producing a single $\hat{y}$ out of these K+1 regressions. Based again on the successes of ensemble methods, the approach taken here is via model averaging:

$$\hat{y} \equiv \bar{\hat{y}}_k = \frac{1}{K}\sum_{k=1}^{K} \hat{\beta}_k X_k + \mathbf{Z}\hat{\mathbf{\Gamma}} \tag{2}$$

That is, the predicted $\hat{y}$ is simply the average of each predicted $\hat{y}_k$ for each of the K one-at-a-time regressions, plus $\mathbf{Z}\hat{\mathbf{\Gamma}}$, where $\hat{\mathbf{\Gamma}}$ is vector of coefficients from regressing $y$ on the controls $\mathbf{Z}$ alone. The most accurate model is the one that best fits $\hat{y}$ to $y$ out of sample, where the metric is the simple root mean square error (RMSE) $= \sqrt{1/N \sum (y_i - \hat{y}_i)^2}$. Again, the out-of-sample aspect is essential, because with numerous independent variables, it may be quite easy to fit the model in-sample, but we can see the estimated coefficients are mistaken when, turning out-of-sample, the over-fit model predicts terribly – as is the case when we regress vote intention against 359 ad count variables at once.[27]

Each out-of-sample test is repeated 500 times, where for each run, the in-sample is a randomly selected 95% of the 859 observations, while the out-of-sample testing group is the remaining 5%.

[26] See Friedman, Hastie & Tibshirani (2009), Ch. 7, for an extensive discussion of the problems in estimating in-sample versus out-of-sample error rates.

[27] And indeed, there probably is good reason to suspect that many hundreds of papers with borderline-significant results would fail this more stringent test.

The essential point is that, if the ads offer no advantage over the controls alone in predicting (ie, explaining the variance in) vote intention, then adding those variables to a model will in fact worsen the predictive accuracy than just employing the controls alone. The higher the proportion of out-of-sample runs where the ads+controls model out-performs the controls-only model,[28] the better that model of advertising measures the true effects of the ads.

The results of these out-of-sample tests can be seen in Figure 1. Each line shows the proportion of out-of-sample runs where the specified model produces a lower RMSE than the controls-only model (longer is better). By binomial chance, we would expect scores due only to sampling variation to fall between 0.46 and 0.54 about 95% of the time; anything outside of these bounds is likely to have over- (or under-)performed due to being genuinely better (or worse) than the baseline.[29] To give a sense of the magnitude of these effects we can remove a control variable: Party ID is of course the most significant one, and line 2 shows that without that variable the model (ie, all the other controls) does worse than the all-controls model on 100% of the out-of-sample runs. On the other hand, removing any one of the other controls (all of which are significant by traditional OLS measures) doesn't hugely damage prediction: on average, removing a non-Party-ID control (line 3) means the new model still beats the baseline 40% of the time – significantly worse, but not as damaging as removing Party ID.

[Figure 1 about here]

But the important question is how well the various competing models of advertising work, shown in the second section of Figure 1. If we regress vote intention on the various category

[28] Which by themselves account for more than 60% of the variance in the dependent variable.

[29] A simple means-difference test for the 500 RMSE's from each method finds significant differences between all of them, although this approach is a bit less reliable inasmuch as one could always run more and more out-of-sample tests until even the smallest accidental difference due to quirks in the data became significant.

variables[30] either individually (not shown) or all together in a single regression (line 4), including the ad-related variables worsens out-of-sample prediction; either the ads have no effect on vote intention, or modeling their effects using a few hand-coded themes and strategies is ineffective. In line 5, we unsurprisingly see that using all 359 individual ad count variables at once in a single multivariate regression is much worse than ignoring the ads altogether, presumably due to extreme over-fitting.

Lines 6 and 7 show that, conversely, looking for a few latent dimensions (PCA)[31] or a few key ads (LASSO) does just barely improve prediction over the controls alone. This shows that the ad broadcast variation does have some correlation with vote intention over and beyond what we see in the controls – although this is only barely statistically significant (the green line), and it still leaves us with latent dimensions or a small handful of ads that are very hard to generalize to a larger theory of persuasion, and even harder to test such generalizations. By contrast, in line 8 we see the results of the one-at-a-time procedure, which suggests that the ads do indeed have a significant effect when measured correctly. More importantly, this appears to be the only successful way to measure those effects: assume that every ad has an effect, however small, estimate those effects separately, and average them all to make a prediction. Line 9 is a slight variation on this: instead of using the individual estimated effects, we shrink them back toward the overall mean effect in proportion to the certainty of the OLS estimate for each of those effects.[32] But although

[30] A category variable counts the total number of broadcasts of ads of that category in that region during that week. CMAG categories include topics such as: whether the ad is an attack or self-promotion; whether it cites evidence, promotes action, is a rebuttal, includes images of the candidate, is policy or personal; and issues such as the economy, women, education, health, the poor, the elderly, crime, morals, or drugs. For estimation, each category variable must be split into two based on the sponsoring party, since we expect them to have opposite effects on vote intention.

[31] This shows the results for 6 components (chosen using scree plots), but results are similar for somewhat fewer or more.

[32] The shrinkage employed here is Bayesian: if we assume the true coefficients are distributed $\boldsymbol{\beta} \sim \mathcal{N}(\bar{\beta}, \tau^2)$ and each measured coefficient is drawn from $\hat{\beta}_k \sim \mathcal{N}(\beta_k, \sigma_k^2)$, then the shrunk coefficients are

$$\hat{\beta}'_k = \frac{\sigma_k^2 \bar{\beta} + \tau^2 \beta_k}{\sigma_k^2 + \tau^2} \approx \frac{\hat{\sigma}_k^2 \bar{\hat{\beta}} + \tau^2 \hat{\beta}_k}{\hat{\sigma}_k^2 + \tau^2}$$

these are presumably individually better estimates of each given ad's effect (reducing spuriously high purported effects), as can be seen on line 9 of Figure 1, the shrunk coefficients predict out-of-sample only slightly better than the unshrunk coefficients (line 8), so the simpler direct estimates will be used hereafter. (For the third section of Figure 1, see the Stage 2 results below.)

### 3.2 Stage One results: aggregate advertising effects

We have established that there is indeed a relationship between advertising and vote intention, but this small effect appears only detectable with the correct structural model of how these effects work: Advertising appears to persuade not via a latent dimension or theme shared by the ads, nor via a few effective ads amidst a majority of ineffective ones, but via a large set of small effects that all the ads participate in to some degree.

At this point we have a set of 359 ads, each with an associated estimated effect, but before we turn to Stage Two, where we look at the contents of those ads, we can already glean some insight into the magnitude and general characteristics of their aggregate effects. To get a sense of how the two parties' success differs, Figure 2 shows a histogram of estimated ad effects grouped by sponsoring party (or that party's supporters). There are two notable things to be seen here: First, the vast majority have very small effects. Second, we can immediately see that many Democratic-sponsored ads appear to have pro-Republican effects, and vice versa. Clearly the campaigns are not nearly as effective as they might believe, and indeed are little better than random in most of their efforts. Although the outliers may be there due more to noise than real effects, Figure 2

where the latter quantity contains the estimated values: $\bar{\bar{\beta}}$ is the mean of the measured coefficients and $\hat{\sigma}_k^2$ is the OLS-derived variance on the $k$th coefficient. The only unknown quantity is $\tau^2$, the true variance, which is estimated by cross-validation. The practical result of this is that ads with apparently high effects, but also high variance (usually due to having been broadcast in only a few locations) are shrunk back, leaving something at least closer to the real effects. This two-step approach is mathematically very similar to the shrinkage and borrowed power of single-step multilevel methods.

strongly suggests that many ads are having opposite their intended effects.[33] Indeed, dropping the very low-effect or backfiring ads nevertheless worsens out-of-sample performance: even those tiny, individually noisily-measured effects are still significant in aggregate.[34]

[Figure 2 about here]

One way to get a sense of the overall effect of advertising is by looking at first differences: shifting party ID by one standard deviation around its mean (while all other variables are held at their means) produces about a 12 point shift in vote share, while most of the demographic variables produce a 1-2 point shift. Shifting the overall level of advertising (from below-actual to above-actual, without changing its distribution), on the other hand, produces about a 2-4 point shift in the pro-Kerry direction,[35] showing that the total effect of adversing was small but potentially comparable in size to other established demographic effects. But a better way to understand the magnitude of the potential advertising effect is to consider a more realistic hypothetical case: Say a candidate had the luxury of going back and doing her campaign over, using the same ads, but selecting only the better ones based on what she learned here. How much could she improve her outcome? Figure 3 shows the estimated outcome were one candidate to buy only the most effective ads,[36] those above some cutoff. Were either candidate to have bought no ads at all, the outcome would be virtually unchanged, since both sets of ads have nearly 0 mean effect (not shown in figure). But had the Kerry campaign bought exactly as many ads, but only those with a coefficient greater than 0 (ie, only those estimated to have a pro-Democratic effect), his outcome would have

[33] A t-test on these distributions finds that the two groups in fact have means that are not significantly different, although a more discerning non-parametric Kolmogorov-Smirnov test shows that the distributions are in fact slightly different.

[34] Although dropping low-effect ads after shrinkage does improve out-of-sample performance slightly, suggesting potential avenues for improving on LASSO by making its estimates univariate rather than multivariate.

[35] Since there are so many variables, this is done by constructing a single average variable, $v = \frac{1}{j}\Sigma_1^j \hat{\beta}_k X_k$, and shifting it.

[36] That is, the candidate runs the exact same number of broadcasts, only distributed differently among their ads.

been significantly, if slightly, improved (line 4). Were he to have bought only the top 25% best performers (coefficients $> 0.001$, line 5), he would have had a significant chance of winning the election. And were he to have broadcast only the top 5% (line 6), the outcome would have been dominantly in his favor, with even the lower bound of the 95% confidence interval above 50% of the two-party vote share. Interestingly, although the pattern is similar for Bush, the effects are much weaker, with only a slight improvement even from choosing the highest-performing ads.[37] Thus, although both actual campaigns were about equally ineffectual, this is not because advertising has no effects or even because the two sides cancelled out (Bartels 1992, Bartels 2006). In fact, many more strong pro-Democrat ads were produced – if only the campaign had better known which ones to emphasize.[38]

[Figure 3 about here]

Having established the complex causal structure and substantive significance of advertising's effects, the obvious next question is *why* some ads are more effective than others. One approach to answering this question is to return to categorization: is the set of negative Democratic-sponsored ads more effective than the non-negative Democratic-sponsored ads, for instance? We might do the same for all of the various hypothesis-driven categories discussed earlier, comparing mean effects between groups to discover effective strategies. WiscAds/CMAG have provided many such

[37] Kaid (2005) shows somewhat similar results for the 2004 campaign, finding that advertising overall had a largely pro-Kerry effects, including even many Bush ads; they find similar pro-Democratic results for 2008, interestingly (Lee Kaid, Fernandes & Painter 2011). Compare this with Shaw (2006), who finds little to no advertising effect in 2004 – although this may be due to endogeneity issues.

[38] Presumably these strategic outcomes are still exaggerated by the fact that many of the largest estimated ad effects are still due, in part, to measurement error on top of their true effect size. Shrinkage was designed to mitigate this, but of course it cannot have eliminated it, so these estimates due to strategic buying should perhaps be taken as upper bounds. However, out-of-sample testing shows that even eliminating ads with small effects, or opposite their intended effects, or even high effects, all worsens out-of-sample prediction, showing that these estimated coefficients, however noisy, are not due to sampling error. Further evidence that highest-effect ads are not there due solely to error is simply that the most pro-Democratic effect ads were in fact sponsored by Democrats, and similarly for most of the pro-Republican effect ads, and these differences are even more distinctive when we turn to the text analysis that reveals the underlying strategies.

categorizations ready-made for this test. But given how poorly these category variables performed in the out-of-sample testing, it is no surprise to discover that almost all of these tests find no significant differences.[39] Once again, either the categories are poorly assigned, or these are not the themes that best explain the variance in ad effects. Instead, we will have to turn to a bottom-up analysis of the raw contents of those ads, their text.

## 4 Stage Two: Inferring causation from text

There are a wide variety of approaches to the statistical analysis of text, but most begin by severely summarizing the set of documents, reducing each to a simple count of words, and throwing away the rich information contained in word order, grammar, and so forth. The remaining "bag of words" is often represented as a matrix of word counts or proportions, here denoted $\Psi_{j,i}$, where there are J = 359 unique ads, I = 1000 unique words,[40] and each entry in the matrix is the proportion of word $i$ in document $j$. But what makes us think there is a discernible connection between the text of ads and their effects, especially after so crudely simplifying those texts? The fundamental test is again via out-of-sample prediction: can we find a way to algorithmically predict the effect of an ad based only on its textual similarity of ads whose effects we already know? If so, then we know that the text is intimately connected to the effect, even at this simplified remove. Once that has been established, we can then seek to find exactly which words are more systematically connected to the highest pro-Democrat and pro-Republican effect ads.

There are two general approaches to text analysis: supervised, and unsupervised. In unsu-

[39] Regressing vote intention on each CMAG category (or all of them together) or doing difference-of-means tests on the $\beta_j$ coefficients, finds that the most significant categories are economic, personal attacks, and rebuttals. Although these results are suggestively in accord with the word-based analysis below, significance levels vary by method, effects often have the wrong sign (pro-Republican effects by Democratic-sponsored attack ads, eg), and in any case we should expect a few to appear significant just by chance when evaluating so many variables.

[40] These are the 1000 most frequent words in the corpus, not including "stop words" like "the," "of," etc.

pervised approaches, one simply takes $\Psi_{j,i}$ and finds latent dimensions, clusters, or topics, and then looks for connections between those sui generis latent qualities and the empirical quantities of interest.[41] Unsurprisingly, tests show that this approach works poorly in this case, since we have a very specific set of numbers associated with each ad – its one-at-a-time measured effect – which do not correlate well with, say, the principal components of $\Psi_{j,i}$. Supervised approaches, on the other hand, employ exogenously supplied categories or scalings in order to predict the characteristics of new documents, exactly what is needed here.[42] To illustrate the power of supervised scaling, we can attempt to replicate a few of the expert WiscAds/CMAG ad classifications using only automated methods. For instance, three important categories are whether an ad is an attack or a self-promotion; whether it is personal vs policy-oriented; and whether it was sponsored by pro-Democrat or pro-Republican groups.[43] In each case, given a set of expert-classified ads and the text of unclassified ads, those unknown ads can be "correctly" classified (ie, matching the human classification) over 80% of the time using methods similar to those used below. Furthermore, the words most associated with each category are quite interpretable: *dollars, tax, iraq* for the attack ads, versus *country, believe, healthcare* for promote ads; *war, country, vietnam* for the personal, versus *tax, plan, health* for the policy ads; *health, plan, companies* for Democratic ads, versus *war, congress, terrorists* for the Republican ads. (See Appendix C for methodological details.) Although there are good reasons to doubt the utility of the CMAG categories, the fact that this automated categorization can be both predictively useful and quite interpretable illustrates the power of supervised scaling.

The essential question is not identifying the themes and topics underlying everything that was

[41] For unsupervised categorization in political science, see Grimmer & King (2011) and Grimmer (2010). For unsupervised scaling, see Monroe & Maeda (2004) and Slapin & Proksch (2008).

[42] For supervised scaling in political science, see Laver, Benoit & Garry (2003).

[43] These particular categories were chosen for their importance in studies such as Jacobs & Shapiro (2000), Popkin (1991), and Brader (2006).

said, but identifying what makes some ads more persuasive than others. To establish this, we must show first that we can predict ad effects based only on their text, and then use those same methods to identify the words and themes underlying specifically the most effective strategies. The first test is to imagine that we do not know the persuasive effect of an ad and then to see how well we can guess its effect, based purely on its textual contents ($\Psi_{j,}$), the contents of the other ads ($\Psi_{\text{-}j,}$), and the "known" effects of the other ads. That is, we are using $\Psi_{j,i}$ and the set of (one-at-a-time estimated) $\beta_{\text{-}j}$ effects for each "known" ad to predict $\hat{\beta}_j$ for some "unknown" ad. There are a two different ways to measure the success of a prediction method: first, simply by taking the RMSE between the out-of-sample, one-at-a-time-estimated $\beta_j$ coefficients, and the predicted coefficients $\hat{\beta}_j$. But although this would allow us to say which method appears to work best at matching $\hat{\beta}_j$ to $\beta_j$, the question remains whether any of those sets of $\hat{\beta}_j$ are really accurate enough to be useful. And the measure of that is not to take the RMSE between $\hat{\beta}_j$ and $\beta_j$, but to use the $\hat{\beta}_j$ values in the original one-at-a-time system, plugging them back in the original out-of-sample-test to determine whether the text-predicted effects are sufficiently accurate to predict vote intention better than the controls alone. This is a much tougher and more substantively relevant test than is common in computer science, requiring not just that we best match predicted categories or scalings to established ones, but that these matches be substantively useful. Glancing ahead, what we will discover is that some methods are better at the easy task (matching $\hat{\beta}_j$ to $\beta_j$) but others are better at the real-world task, and best of all is once again the ensemble approach, combining three very distinct theoretical approaches.

As before, a number of different approaches are developed (or adapted) here, not for the sake of methodological completeness, but because each approach depends on, and tells us something about, a different causal mechanism connecting the text of ads to their effects. The first assumes that every word, or the concept it evokes, has a separate effect that directly aggregate to the overall effect of the ad. This is essentially the one-at-a-time regression method: instead of regressing vote intention on broadcast counts, we regress ad effects on word counts, and predict unknown effects in the same way that vote intention was predicted before (see Appendix C). The second approach assumes that there is a shared space (like an ideological space) in which every word and ad can be positioned, and the effect of an ad reflects its position in this space; the effect of an unmeasured ad is estimated by taking its position in this space and comparing it to nearby ads whose effects are already known. This comparison can be by taking the distance-weighted average of all ads, or by averaging the nearest K neighbors, or by averaging only those within some fixed cutoff distance; these variants assume increasing degrees of idiosyncrasy in ads, where only ads very similar to an unknown ad are relevant to predicting its effect (see Appendix C). The third approach is Bayesian, considering each ad to define a type or category, where an unknown ad's effect can be approximated by estimating the expected value of its type, given our knowledge of each other ad's type (see Appendix C). As we will see, these approaches have different strengths, which in turn illuminate how their texts influence their effects.

Table 1 shows the correlations between the $\hat{\beta}_j$ coefficients estimated using each of these three methods and the original $\beta_j$ coefficients. The apparent result is that the one-at-a-time word regression works very well, correlating highly with the original coefficients; the Bayesian approach works less well, but still significantly; while the spatial approaches have no significant correlation.[44] The latter's poor performance is due to the fact that usually only a very few ads are employed in estimating the unknown ad – the optimal cut-off includes only a few ads, the best K of KNN is

[44] Although using a Spearman rank correlation test does find significant, if small, correlation between the original coefficients and the hard cutoff and KNN results.

2, and even the weighting tends to weight distant ads very little.[45] But as the inter-correlations show, these results are not spurious: they correlate with the word regression and Bayesian output significantly, just not with the original coefficients.

[Table 1 about here]

But again, what we are fundamentally concerned with is measuring, predicting, and explaining vote intention; the estimated effect coefficients are only a means to this end. If we were to claim that we could predict the effectiveness of an ad based only on this comparison with original coefficients, a skeptic might reasonably wonder whether the correlation with direct measurement was close enough to be of any practical use. The real test is whether we can use text to predict ad effects sufficiently well that, when these text-estimated ads effects are employed in out-of-sample prediction of *vote intention*, they improve at all on the controls-only baseline. The results of this more substantive test are presented in the third section of Figure 1 (lines 10-15). As before, each line shows the proportion of out-of-sample runs where the ads model out-performs the controls-only baseline. But in these cases, the coefficients are derived from the specified technique, predicted using only the textual similarity between the out-of-sample ad and the in-sample ads. The surprising result is that the one-at-a-time and Bayesian techniques (lines 10-11) – which best matched the original coefficients – do least well at predicting vote intention, while the spatial approaches all do significantly better than chance (lines 12-14). This difference in “real world” performances appears to be because, while the regression, Bayes, and (to a lesser degree) weighting methods all utilize every known ad to predict the effect of an unknown ad, the KNN and hard-cutoff methods generally use only the most similar ads.

[45] Unlike the Bayesian and one-at-a-time approaches, the spatial methods all have one or two parameters to fit, which were selected by maximizing out-of-sample predictions for a subsample of the data.

In addition to constituting a potentially powerful tool for campaigns to test new ads, these results provide an important insight into how ads work: not through broad effects that are present in all ads to greater or lesser degrees, but through a wide array of different mechanisms that are not quite idiosyncratic to each ad, but are best captured through looking at only a few related ads. In conjunction with the results from Stage One, this suggests an exceedingly complex set of causal pathways connecting the content of ads to their effects: we must look at every ad, because even the smallest have an effect; nor can we simplify the themes and strategies underlying the effects of those ads, because they too are numerous and idiosyncratic (though not quite so idiosyncratic as to defy all efforts at systemization, as we will see). And indeed, things are even more complex than that: line 15 shows that, while the spatial approaches work best, even better is an ensemble of all of the disparate methods. The coefficients generated by averaging all of the text-based predictions work better than even the best of the spatial methods, suggesting that in addition to the idiosyncratic effects, there may also be some weaker, more general effects due to the words, positions, or themes of even fairly unrelated ads.[46]

## 4.1 Stage Two results: scaling words to infer effective strategies

Having determined that there is a systematic relationship between the texts of ads and their effects, as social scientists we still want to know *why*: what characterizes the most successful ads?[47] By determining which words are most associated with the most successful pro-Democrat

[46] These "predictions" have all been out-of-sample, but using the pooled dataset. True prediction, of course, does not work this way, nor would a real campaign. One can also test these methods truly predicting forward in time, by estimating effects using the weeks prior to some time $t$, then predicting the effects of new ads in week $t+1$ and testing models as before. This is more like how a real campaign might use these tools, in order to test new ads for their potential efficacy in the week ahead. There are not enough weeks to establish true statistical significance for this test, but for the 9 weeks leading up to the election, for two thirds, the text-based prediction of vote intention is more accurate than the controls alone, strongly suggesting that these approaches work just as well at true prediction as they do for pooled out-of-sample testing.

[47] Even practical-minded campaigns would like to be able not just to evaluate the likely effectiveness of new scripts, but to know what words and concepts to build newly-written ads around.

and pro-Republican ads, we can identify which persuasive strategies are most effective in this election out of the vast array of possibilities discussed in the introduction. As we will see, the text analysis provides a unique insight into a complex mix of successful strategies, which combine policy positioning, issue ownership, negativity, fear, and promise in combinations that distinctly differ between the two parties.

Just as the one-at-a-time method assigns each ad a coefficient measuring its pro-Democratic or pro-Republican effect, we wish to score each of the thousand words in the corpus according to the overall persuasive effect of the ads in which it appears. For each of the text-based prediction methods above, a cognate word-scaling method can be devised, each of which again has a corresponding causal mechanism underlying it: determining the independent effect of each word through its correlation with successful ads (the one-at-a-time approach); fitting words and ads into a shared, low-dimensional ideology-like space (the spatial approach); and treating the words in an ad as random variables drawn from distributions associated with the effectiveness of an ad (the Bayesian approach). (See Appendix C for more details.) In this last stage, however, there is no direct way to subsequently test these word scoring models to determine the best; rather, the test was earlier, when these methods were shown to predict effects well. And since an ensemble of these methods worked best for prediction, a similar approach is taken here, combining the word scalings into a single, unified scaling. As before, since each method is likely to produce many words that score highly by mere chance, combining multiple methods reduces the chance that we will end up with words that score highly due merely to a quirk in one method or another. And finally, here we will also need error estimates for each word's score, both for combining the scalings, and for the interpretation stage, where we will want to examine only those words whose scores we are most confident of.[48]

The best way to combine the scalings from these three methods is via Bayesian model averaging (BMA) (Madigan, Raftery, Volinsky & Hoeting 1996), which essentially takes the weighted mean of each word's score from each model, weighted by the confidence level of that score.[49] Once we have our best overall scaling of the words, though, how are we to interpret it? Once again, we have a surfeit of numbers – not 359 as in stage one, but now 1,000. And even after averaging, many of the highest-scoring pro-Democratic and pro-Republican words might be there solely by chance. But having calculated errors for each score, we can reduce the effect of spurious scores by focusing only on those we are most confident in, such as those with a p-value less than 0.05.[50] Figure 4 shows a plot of each word word according to its BMA score (x axis), and how many times any ad with that word in it was broadcast (y axis). Words in bold are those with scores at least two standard deviations from zero. (See Appendix C for Figure 6, showing only the top words along with their scores and 95% confidence intervals.) As can be seen in Figure 4, the vast majority of words are both infrequent and have little effect. Looking only at the distribution of the most significant words (in bold), we can already see one asymmetry between the two sides: on the Democratic side, we see more low-impact, high-frequency words, suggesting strategies that are successful though the gradual accumulation of many small nudges; whereas the Republican side we have infrequent but

[48] Error estimates for the one-at-a-time approach are simply the OLS standard errors, whereas for the other two approaches, errors are calculated via bootstrapping; see Appendix C.

[49] The BMA calculation resembles the Bayesian shrinkage calculation in note 32, with a corresponding calculation for updating the posterior standard error. However, there is one hitch in this case, a weakness of the BMA method when combining very disparate models. The Bayesian scaling systematically produces much smaller errors than the other two (even after normalization), and this over-confidence leads to the other outputs being overwhelmed during the averaging process. Furthermore, although the standard errors for the word scores are moderately correlated for the spatial and regression methods, the Bayesian errors are uncorrelated with the others. Since the Bayesian scores are correlated with the spatial set at 0.95, the best solution is simply to drop the Bayesian scoring, and combine the other two for the best overall scaling.

[50] These p values can also be adjusted for multiple testing bias, since many low p-values may appear only by chance due to testing so many variables. Applying the conservative Bonferroni step-down correction to avoid false positives results in only a quarter as many significant words, but this is likely too stringent for our purposes; seeking a low False Discovery Rate (using the Benjamini and Hochberg adjustment) produces a word list not very dissimilar from the unadjusted 0.05 threshold (Benjamini & Hochberg 1995, Dudoit & van der Laan 2008).

higher effects, suggesting success through micro-targeting rather than blanketing.[51]

[Figure 4 about here]

Turning to the meanings of the statistically significant words, we can immediately see a further asymmetry between the two sides.[52] On the Democratic side, the great majority of the words concern healthcare and prescription drugs. This appears to have worked slowly but surely through repeated and small effects, though a few strategies produced higher effects, mainly the more aggressive ads[53] that emphasize tough negotiation and importing drugs from Canada, even at the risk of going to jail ("jail" also turns up in ads tying Bush to Enron). On the Republican side, we see a variety of low-frequency, higher-impact strategies: the infamous Swift Boat Veterans for Truth, abortion, Osama bin Laden, the deficit, Kerry's voting record in the Senate,[54] the libertarian candidate Michael Badnarik, and more generally, evocations of "worry" and "concern." But unlike on the Democratic side, a few of these pro-Republican words actually reflect Democratic-sponsored ads: the ads mentioning Osama bin Laden, and some of the ads mentioning the Swift Boats and the deficit. We cannot determine from these data whether this reflects genuinely backfiring ads or – as is likely the case with the Swift Boat debate – ads which were run by Democratic supporters ineffectually rebutting Republican attacks in the same locations as those attack ads. In any case, these war-related ads were not successful for Democrats either as attacks or rebuttals, and may

[51] See Hillygus & Shields (2008) for similar results for the 2004 election, where the Bush campaign appears to more aggressively and successfully micro-target independents and wavering Democrats, whereas the Kerry campaign appears to ineffectually "microshield" its own partisans.

[52] Obviously these meanings cannot be interpreted without knowledge of, and reference to, the context of the ads and political environment at the time. But this too is analogous to, for instance, genotyping a disease: we learn all the genes associated with a disease, but to understand the causal mechanisms, we must bring to bear our knowledge of the system in which those genes operate.

[53] Many run by MoveOn, a word scoring just below the $2\sigma$ threshold. This and the Swift Boat ads emphasize the importance of independent groups in actually persuading voters in this campaign.

[54] While most of the "vote" words on the Republican side relate to this, "voter" on the Democratic side mainly reflects the various pro-Democratic sponsoring groups with "voter" in their names. This illustrates, incidentally, the importance of not "stemming" – reducing all cognate words to a single stem like "vot*".

indeed have weakened their candidate by emphasizing issues that did not play to his strengths.

Returning to the higher-level persuasive strategies discussed in the introduction, clearly many of the successful strategies on both sides involve framing and issue ownership, shifting the terms of debate towards each candidate's strengths. And reflecting the importance of the more fluid image of the challenger, both sides appear to focus more on Kerry, whose perception by the public may be more malleable than that of the incumbent Bush. Even the presence of Michael Badnarik – who helped the Republicans in 2004 by convincing dissatisfied Bush voters to vote for the libertarian rather than Kerry – is more about diminishing votes for Kerry than increasing them for Bush. There is very little issue convergence, and what little there is (the deficit, Vietnam, the war on terrorism) seems to mainly help the Republican side, illustrating the dangers of this tactic.

More generally, the two sides differ in the sorts of issues their successful ads emphasize: There is more foreign policy on the right, more domestic policy on the left. The are more retrospective issues on the right (Kerry's voting and Vietnam histories), while more prospective issues on the left (future drug policy). And more broadly, there are more attacks and negative emotions ("worry", "concern") on the right, versus less emotional, more policy-oriented proposals on the left[55] – where even the more aggressive, higher-scoring words involve "negotiation." There is good reason to think that these asymmetries are not accidental: each side is successful when it plays to its strengths on a variety of dimensions – policy, emotional, temporal, physical, degree of micro-targeting – rather than converging on a single set of strategies. And where they do converge, one side appears to do the worse for it.

[55] This policy/emotion dichotomy resembles the results in Westen (2007), and many earlier studies. Geer (2006) also finds that positive ads are often more prospective, while negative ones are more retrospective; but whether negative Bush ads like the Swift Boat interviews, or (debatably) positive Kerry ads about health policy are more informative is difficult to say. Ridout & Franz (2011) find that negative ads were ineffectual in 2004, as was the emotional content of ads except in slightly helping Kerry, but these results may reflect once again measurement difficulties with using the WiscAds/CMAG categories, or with categorization more generally.

An essential point here is that these are not characterizations of everything each side attempted; while it is interesting to know every last thing the campaigns thought to throw against the wall, much more important is understanding what was actually persuasive. But knowing this for 2004 is still just a start: to generalize these results further will require looking at other election years; at regional variations; at congressional and local races; and at interactions with individual-level characteristics. And it must be stressed again that these word-based results are by no means self-explanatory, requiring extensive knowledge of the issues and contexts at play in a given election. But this expertise is no more than goes into any categorization of an ad or speech as negative, personal, policy-oriented, and so on.[56]

As we have seen, beginning with one or two simple theories of persuasion and categorizing ads accordingly works quite poorly to measure the myriad and asymmetrical strategies simultaneously at play in a real campaign. Almost as unsuccessful were the established simplifying approaches, such as searching for an underlying dimension of persuasion (PCA) or a few killer ads doing most of the work (LASSO). Treating all of the ads as contributing in small and varying ways, and measuring those effects using one-at-a-time regression, was the only approach that successfully measured the persuasive effect of advertising – and found that, contrary to prevailing wisdom, this effect was substantial and could be much more so with better-informed campaigns. To understand why some ads were more effective than others requires a text-based approach that similarly embraces the underlying complexity of real-world persuasion. Automated text analysis, although it begins with the simplification of ads to word counts, nevertheless shows that with the proper model (reflecting

[56] Note that one could potentially test these word scores by constructing variables which, say, count the number of high-scoring words in every ad run in a given region-week. But tests show that these bottom-up category variables work little better than the WiscAds/CMAG ones. This is presumably because there is no firm line between words that "work" and words that don't, just a series of gradations; and while one could create more sophisticated variables that reflected the weighted average of every word that appears (weighted by that word's BMA score), this would essentially just be reconstructing the original ad scores, and offers no additional insight into the accuracy of the word scores themselves.

the diversity and idiosyncrasy of ads), deep connections can be found between texts and effects, allowing campaigns to predict the effects of ads before they are run. And having proven this approach, turning it to word scaling allows us to easily read off and potentially explain those effects via numerous but readily interpretable sets of words and concepts. Most importantly, the steps developed here constitute a more general bottom-up procedure for measuring, validating, and understanding the persuasive effects of linguistic data, and should be applicable not just to future work in campaign analysis, but to a wide variety of domains where actors attempt to influence each other with speech.

## Appendix A: Omitted variable bias?

One worry is that the one-at-a-time approach may be hindered by omitted variable bias, since each univariate regression is omitting hundreds of variables.

Specifically, if we know that the true model is:

$$Y = X\beta + Z\delta + \epsilon \tag{3}$$

where X and Z are two sets of independent variables that together constitute Y, and we instead regress Y on X alone, we get a $\hat{\beta}$:

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'Y \\
&= (X'X)^{-1}X'(X\beta + Z\delta + \epsilon) \\
&= \beta + (X'X)^{-1}X'(Z\delta) + (X'X)^{-1}X'\epsilon \\
&= \beta + (X'X)^{-1}X'(Z\delta) \qquad (4)
\end{aligned}$$

assuming iid errors. Thus our estimate for $\hat{\beta}$ has a bias due to regressing Y on only the X variables, equal to the last term in (4). In general, one can say little about the magnitude or direction of this error (Greene & Zhang 2003). But say we believe that all the ad variables in the set of 359 should be included, but regress Y on only one of them at a time. Then each $\hat{\beta}$ will be biased by the omission of all the other ad variables, but in a predictable way. Specifically, for the $j = 359$ individual ad regressions we get for the ad coefficients (omitting the controls):

$$\begin{aligned}
\hat{\beta}_1 &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'X_3\beta_3 + \ldots \\
\hat{\beta}_2 &= (X_2'X_2)^{-1}X_2'X_1\beta_1 + \beta_2 + (X_2'X_2)^{-1}X_2'X_3\beta_3 + \ldots \\
\hat{\beta}_3 &= \ldots \\
&\ddots \qquad (5)
\end{aligned}$$

or

$$\hat{\beta}_i = \sum_{k=1}^{j} \hat{\psi}_{i,k}\beta_k \tag{6}$$

where $\hat{\psi}_{i,k}$ is the coefficient from regressing $X_i$ on $X_k$. So if we assume that the $j$ variables are everything (a strong assumption, of course), then we can solve for the true $\beta_k$ coefficients: since we can calculate all the $\hat{\beta}_i$ and $\hat{\psi}_{i,k}$ coefficients, solving for the $\beta_k$s is just a matter of $j$ equations and $j$ unknowns.

At this point, certain computational practicalities obtrude: with 359 simultaneous linear equations and a matrix of coefficients with a determinant on the order of $10^{-500}$, this system cannot be solved analytically. Ironically, the best solution is found (say, by Mathematica or Matlab) by simply carrying out the multivariate regression, regressing the vector of $\hat{\beta}_i$ on $j$ $\hat{\psi}_{i,k}$ vectors. Using that method, we can find a fairly close (if approximate) solution to the system, resulting in a set of "true" $\beta_k$ values.

Returning to the original question now, does this "correction" for omitted variable bias help? The answer is no. When we test this procedure out of sample, only 63% of the runs do better than the controls alone. This shows that the procedure is not junk, since as we have seen, most meddling with the coefficient values will produce vastly lower OOS percentages. In fact, the "corrected" coefficients are usually quite different from the original one-at-a-time ones, yet they still work fairly well. It's just that regressing each ad one at a time and averaging the final prediction works better, regardless of whatever omitted variable bias that may introduce. In fact, it may be that "correcting" the coefficients merely reintroduces some of the same problems of over-fitting that the one-at-a-time method sought to avoid. The power of the one-at-a-time method derives precisely from omitting all the other variables: when there are so many variables, the bias due to omission is greatly exceeded by the gains made in avoiding over-fitting.

## Appendix B: Monte-Carlo validation of one-at-a-time

The success of the one-at-a-time method, as well as its limitations, can be seen via a simple Monte Carlo simulation. A linear model is constructed as $y = \mathbf{X}\boldsymbol{\beta} + \epsilon$,[57] and each of four methods are tested by fitting them on a subsample of the data and then predicting $\hat{y}$ out-of-sample. Figure 5 shows the out-of-sample predictive accuracy for multiple regression, LASSO, one-a-time regression, and the simple (in-sample) mean $\bar{y}$. The metric of predictive accuracy is the root mean square error (RMSE) $= \sqrt{1/N \sum (y_i - \hat{y}_i)^2}$, as it is throughout this paper. As the noise level $\epsilon$ is gradually increased in generating $y$, the ability of each of the models to predict $y$ goes down. For each set of tests (each repeated 10,000 times) the noise level $\epsilon$ is increased, making it more difficult to predict $\hat{y}$.[58]. Figure 5 shows the predictive accuracy on the y axis (RMSE minus noise), for increasing levels of noise $\epsilon$ (x axis). Lower RMSE scores are better, and the model with the lowest score for each noise level is the one that does the best job inferring $\hat{y}$, and therefore the true coefficients $\boldsymbol{\beta}$. The results are striking: for the lowest levels of noise, multiple regression works best; for higher noise levels, LASSO does the best job; but for yet higher noise, one-at-a-time regression does the best, across a large range of noise levels. At the highest noise level, none of the methods are able to recover any information about the true coefficients, and one is best off eschewing a model and simply guessing $y$ using its in-sample mean. These simulations show that one-at-a-time regression actually does a better job than the popular LASSO method of inferring effects when variables are numerous or noise is high – with the advantage of providing an estimated effect for every variable.

[Figure 5 about here]

[57] Specifically, in this case there are 10 independent variables, which are somewhat correlated with each other ($\sim$ 0.2) and five of the $\beta$ coefficients are set to 0, while the other five are drawn from a uniform distribution for each run. The intercorrelation should give multiple regression a leg up on the one-at-a-time method, while setting five coefficients to 0 should give LASSO a leg up. However, the results are much the same if the variables are not correlated; if the coefficients are all drawn from a uniform distributions; and if there are more or fewer than 10.

[58] Specifically, for each iteration, $\epsilon$ is drawn from a uniform distribution between $-m$ and $m$; for each stage, $m$ increases by one. Figure 5 shows the level $m$ on the x axis, and the RMSE minus $m$ on the y axis.

# Appendix C: Text Methods

## Classifying ads according to key WiscAds/CMAG categories

A procedure similar to the spatial methods employed below which can classify an “unknown” ad as belonging one of $k$ mutually exclusive categories is the nearest centroid approach (Friedman, Hastie & Tibshirani 2009), which simply considers each ad as a point in I=1000-dimensional space (where each dimension corresponds to the proportion for word $i$ in that ad), and classifies the new ad according to whether it is nearest to the center of all classified ads of type A (eg, Democrat-sponsored) or type B (eg, Republican-sponsored). Table 2 shows the classification accuracy for the three category-topics discussed in the text, along with the top words (those the differ most between the opposing centroids of a classification).

[Table 2 about here]

## Predicting the effects of ads using their text

For each method, we wish to estimate the “unknown” effect of ad $\beta_j$ knowing only its text $\mathbf{\Psi}_{j,}$ (that is, row $j$ of $\mathbf{\Psi}$), the text of the other ads $\mathbf{\Psi}_{-j,}$ (other rows in $\mathbf{\Psi}$, though this may be a subsample of $\mathbf{\Psi}$), and the one-at-a-time-measured effects of those other ads, the set $\beta_{-j}$.

The three approaches work as follows:

**One-at-a-time regression**

$$\boldsymbol{\beta} = \gamma_i \mathbf{\Psi}_{,i} \text{ for each word } i : 1...m \tag{7}$$

$$\hat{\beta}_j = 1/m \Sigma_{i=1}^{m} \hat{\gamma}_i \Psi_{j,i} \tag{8}$$

For each word $i$, we generate a regression coefficient $\gamma_i$ by regressing the vector of coefficients from the original one-a-time-procedure (where each ad is an observation) on the vector of word proportions for that word in each of the ads. We then generate the predicted values $\hat{\beta}_j$ by averaging the individual predictions, as before. The one-at-a-time method was devised exactly to suit these situations, where variables are numerous and noisy, regardless of whether those variables are ad broadcast counts, or word proportions.

**Spatial** (variants: cutoff, weighting, KNN)

$$w_{j,k} = 1/\left(1 + \exp(a(\|\Psi_{j,} - \Psi_{k,}\| + b))\right) \tag{9}$$

$$\hat{\beta}_j = \frac{\Sigma_{k=1}^{n} w_{j,k}\beta_k}{\Sigma_{k=1}^{n} w_{j,k}} \tag{10}$$

In this case, we take each ad as a position in word space (a simplex in 1000-dimensional Euclidean space, in this case). Each predicted coefficient $\hat{\beta}_j$ is simply the weighted average of all those ads $\beta_{-j}$ around it, where the weight is determined by the distances of those ads from the "unknown" one. This weighting function can be logistic with parameters $a$ and $b$, which implies that ads may share a low-dimensional space where every ad's position affects every other's. Alternatively, one can use a hard cutoff, where only those ads within a certain distance of the unknown one are included, or only the k nearest neighbors (KNN) are. These variants instead suggest that ad effects are more idiosyncratic, where textually similar ads may have similar effects, but more textually dissimilar ads offer no insight into the effect of an unknown ad. All three variants are tested here.

**Bayesian**

$$\mathrm{P}(\Psi_{j,}|L_k) \propto \Pi_{i=1}^{m} \Psi_{k,i}^{\Psi_{j,i}} \tag{11}$$

$$\hat{\beta}_j = \frac{\Sigma_{k=1}^{n} P(\Psi_{j,}|L_k)\beta_k}{\Sigma_{k=1}^{n} P(\Psi_{j,}|L_k)} \tag{12}$$

Like the spatial approach, this is essentially a weighted mean, but in this case, $\hat{\beta}_j$ is the expectation value given the probabilities that unknown ad $j$ belongs to each class $L_k$, where each class $L_k$ is defined by that ad $k$. This probability $\mathrm{P}(\Psi_{j,}|L_k)$ is calculated as in equation (12) in the "naive" Bayesian fashion, where the probability of a word $i$ occurring in the class defined by ad $k$ is simply $\Psi_{k,i}$, the proportion of that word in that ad, and all word probabilities are taken as independent of each other.

## Word scaling using the three approaches

Adapting each of the prediction methods to provide a score for each word corresponding to its pro-Democratic or pro-Republican effect is a relatively straightforward process:

### One-at-a-time

This simply uses the coefficients assigned to words from equation 7. Conveniently, the regression procedure also provides standard errors directly.

### Spatial

For this approach, I adapt the unsupervised scaling developed by Monroe & Maeda (2004) and Slapin & Proksch (2008), turning it into a supervised method. In the original formulation, inspired by item response theory, words and documents are presumed to have positions in a one-dimensional space where (in the original formulation) we assume words are drawn from documents in proportion to the distance between the word and the document. The modification here is that the documents are presumed to already have positions – their one-at-a-time coefficients – and only the words $i$ will be positioned $(x_i)$ to minimize the following likelihood:

$$\arg\min_{x_i} \Pi_{i,j} (e^{-(x_i - \beta_j)^2} - \Psi_{j,i})^2 \tag{13}$$

That is, we minimize the squared errors between, on the one hand, the distance between $x_i$ and the ad position $\beta_j$, and on the other hand, the proportion of word $i$ in document $j$, $\Psi_{j,i}$. I then employ bootstrapping (on the ad/observation level) to estimate errors.

### Bayesian

The approach here is similar to that in equation 12: we want the ad effect expectation given word $i$. Making the usual naive independence assumptions, this is:

$$\mathbf{E}[\beta|w_i] \approx \frac{\sum_j \beta_j \Psi_{j,i}}{\sum_j \Psi_{j,i}} \tag{14}$$

That is, the word score is the expected coefficient given that word, or the sum over ads of each ad coefficient times the probability of that ad given that word. Once again, errors can be estimated via bootstrapping.

[Figure 6 about here]

# References

Abramowitz, A.I. 1991. “Incumbency, campaign spending, and the decline of competition in US House elections.” *The Journal of Politics* 53(01):34–56.

Alvarez, R.M. 1998. *Information and elections*. Univ of Michigan Pr.

Annest, A., R. Bumgarner, A. Raftery & K.Y. Yeung. 2009. “Iterative bayesian model averaging: A method for the application of survival analysis to high-dimensional microarray data.” *BMC bioinformatics* 10(1):72.

Ansolabehere, S. & A. Gerber. 1994. “The mismeasure of campaign spending: Evidence from the 1990 US House elections.” *Journal of Politics* 56(4):1106–18.

Ansolabehere, S., E.C. Snowberg & J.M. Snyder. 2006. “Television and the incumbency advantage in US elections.” *Legislative Studies Quarterly* 31(4):469–490.

Ansolabehere, S. & S. Iyengar. 1994. “Riding the wave and claiming ownership over issues: The joint effects of advertising and news coverage in campaigns.” *Public Opinion Quarterly* 58(3):335–357.

Ansolabehere, S. & S. Iyengar. 1995. “Going negative: How political advertising shrinks and polarizes the electorate.”.

Arceneaux, K, A.S. Gerber & D.P. Green. 2006. “Comparing experimental and matching methods using a large-scale voter mobilization experiment.” *Political Analysis* 14(1):37–62.

Bartels, L.M. 1992. “The impact of electioneering in the United States.” *Electioneering: A comparative study of continuity and change* pp. 244–277.

Bartels, L.M. 1993. “Messages received: The political impact of media exposure.” *American Political Science Review* pp. 267–285.

Bartels, L.M. 1996. “Uninformed votes: Information effects in presidential elections.” *American Journal of Political Science* pp. 194–230.

Bartels, L.M. 2006. “Priming and persuasion in presidential campaigns.” *Capturing campaign effects* pp. 78–112.

Benjamini, Y. & Y. Hochberg. 1995. “Controlling the false discovery rate: a practical and powerful approach to multiple testing.” *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 289–300.

Benoit, K. & M. Marsh. 2008. “The campaign value of incumbency: A new solution to the puzzle of less effective incumbent spending.” *American Journal of Political Science* 52(4):874–890.

Bøvelstad, H.M., S. Nygård & Ø. Borgan. 2009. “Survival prediction from clinico-genomic models-a comparative study.” *BMC bioinformatics* 10(1):413.

Brader, T. 2006. *Campaigning for hearts and minds: How emotional appeals in political ads work*. University Of Chicago Press.

Brady, H.E., R. Johnston & J. Sides. 2006. *Capturing Campaign Effects.* Univ of Michigan Pr chapter The Study of Political Campaigns.

Brasher, H. 2003. "Capitalizing on contention: Issue agendas in US senate campaigns." *Political Communication,* 20(4):453–471.

Breiman, L. 2001. "Random forests." *Machine learning* 45(1):5–32.

Cameron, A.C. & P.K. Trivedi. 2005. *Microeconometrics: methods and applications.* Cambridge university press.

Campbell, J.E. 2008. *The American campaign: US presidential campaigns and the national vote.* Vol. 6 TAMU Press.

Damore, D.F. 2004. "The dynamics of issue ownership in presidential campaigns." *Political Research Quarterly* 57(3):391–397.

Damore, D.F. 2005. "Issue convergence in presidential campaigns." *Political Behavior* 27(1):71–97.

Davenport, T.C., A.S. Gerber & D.P. Green. 2010. "Field experiments and the study of political behavior." *The Oxford Handbook of American Elections and Political Behavior* .

Dietterich, T. 2000. "Ensemble methods in machine learning." *Multiple classifier systems* pp. 1–15.

Dietterich, T.G. 1997. "Machine-learning research." *AI magazine* 18(4):97.

Druckman, J.N. 2001*a*. "On the limits of framing effects: who can frame?" *Journal of Politics* 63(4):1041–1066.

Druckman, J.N. 2001*b*. "The implications of framing effects for citizen competence." *Political Behavior* 23(3):225–256.

Druckman, J.N. 2004. "Political preference formation: Competition, deliberation, and the (ir) relevance of framing effects." *American Political Science Review* 98(4):671–686.

Dudoit, S. & M.J. van der Laan. 2008. *Multiple testing procedures with applications to genomics.* Springer Verlag.

Erikson, R.S. & T.R. Palfrey. 1998. "Campaign spending and incumbency: An alternative simultaneous equations approach." *Journal of Politics* 60:355–373.

Erikson, R.S. & T.R. Palfrey. 2000. "Equilibria in campaign spending games: Theory and data." *American Political Science Review* pp. 595–609.

Ferejohn, J.A. & J.H. Kuklinski. 1990. *Information and democratic processes.* Univ of Illinois Pr.

Franklin, C.H. 1991. "Eschewing obfuscation? Campaigns and the perception of US Senate incumbents." *The American Political Science Review* pp. 1193–1214.

Franz, M.M., P. Freedman, K. Goldstein & T.N. Ridout. 2008. "Understanding the effect of political advertising on voter turnout: A response to Krasno and Green." *The Journal of Politics* 70(01):262–268.

Franz, M.M. & T.N. Ridout. 2007. "Does political advertising persuade?" *Political Behavior* 29(4):465–491.

Friedman, J., T. Hastie & R. Tibshirani. 2009. *The elements of statistical learning.* Springer Series in Statistics.

Geer, J.G. 2006. *In defense of negativity: Attack ads in presidential campaigns.* University of Chicago Press.

Gelman, A. & G. King. 1993. "Why are American presidential election campaign polls so variable when votes are so predictable?" *British Journal of Political Science* 23(04):409–451.

Gerber, A.S., J.G. Gimpel, D.P. Green & D.R. Shaw. 2011. "How Large and Long-lasting Are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment." *American Political Science Review* 105(01):135–150.

Greene, W.H. & C. Zhang. 2003. *Econometric analysis.* Vol. 5 Prentice hall Upper Saddle River, NJ.

Grimmer, J. 2010. "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases." *Political Analysis* 18(1):1.

Grimmer, J. & G. King. 2011. "A General Purpose Computer-Assisted Clustering Methodology." *Proceedings of the National Academy of Sciences* .

Haibe-Kains, B., C. Desmedt, C. Sotiriou & G. Bontempi. 2008. "A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all?" *Bioinformatics* 24(19):2200–2208.

Hillygus, D.S. & T.G. Shields. 2008. *The persuadable voter: wedge issues in presidential campaigns.* Princeton Univ Pr.

Holbrook, T.M. 1996. *Do campaigns matter?* Vol. 1 Sage Publications, Inc.

Holbrook, T.M. 1999. "Political learning from presidential debates." *Political Behavior* 21(1):67–89.

Huber, G.A. A & K. Arceneaux. 2007. "Identifying the Persuasive Effects of Presidential Advertising." *American Journal of Political Science* 51(4):957–977.

Imai, K. 2005. "Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments." *American Political Science Review* 99(02):283–300.

Iyengar, S. 1994. *Is anyone responsible?: How television frames political issues.* University of Chicago Press.

Iyengar, S. & D.R. Kinder. 1987. *News that matters.* University of Chicago Press Chicago, IL.

Iyengar, S., M.D. Peters & D.R. Kinder. 1982. "Experimental demonstrations of the 'not-so-minimal' consequences of television news programs." *The American political science review* 76(4):848–858.

Iyengar, S. & N.A. Valentino. 2000. “Who says what? Source credibility as a mediator of campaign advertising.” *Elements of reason: Cognition, choice, and the bounds of rationality* pp. 108–129.

Jackson, R.A., J.J. Mondak & R. Huckfeldt. 2009. “Examining the Possible Corrosive Impact of Negative Advertising on Citizens’ Attitudes toward Politics.” *Political Research Quarterly* 62(1):55–69.

Jacobs, L.R. & R.Y. Shapiro. 2000. *Politicians don’t pander: Political manipulation and the loss of democratic responsiveness.* University of Chicago Press.

Jacobson, G.C. 1978. “The effects of campaign spending in congressional elections.” *The American political science review* pp. 469–491.

Johnston, R., M.G. Hagen & K.H. Jamieson. 2004. *The 2000 presidential election and the foundations of party politics.* Cambridge Univ Pr.

Kahn, K.F. & P.J. Kenney. 1999. “Do negative campaigns mobilize or suppress turnout? Clarifying the relationship between negativity and participation.” *American Political Science Review* pp. 877–889.

Kaid, L L. 2005. “Political Advertising in the 2004 Election: Comparison of Traditional Television and Internet Messages.” *American Behavioral Scientist* 49(2):265–278.

Kaplan, N., D.K. Park & T.N. Ridout. 2006. “Dialogue in American political campaigns? An examination of issue convergence in candidate television advertising.” *American Journal of Political Science* 50(3):724–736.

Krasno, J.S. & D.P. Green. 1988. “Preempting quality challengers in House elections.” *Journal of Politics* 50(4):920–36.

Krasno, J.S. & D.P. Green. 2008. “Do Televised Presidential Ads Increase Voter Turnout? Evidence from a Natural Experiment.” *The Journal of Politics* 70(01).

Krosnick, J.A. & D.R. Kinder. 1990. “Altering the foundations of support for the president through priming.” *The American political science review* pp. 497–512.

Krupnikov, Y. 2011. “When Does Negativity Demobilize? Tracing the Conditional Effect of Negative Campaigning on Voter Turnout.” *American Journal of Political Science* .

Lau, R.R., L. Sigelman & I.B. Rovner. 2007. “The effects of negative political campaigns: a meta-analytic reassessment.” *Journal of Politics* 69(4):1176–1209.

Laver, M., K. Benoit & J. Garry. 2003. “Extracting policy positions from political texts using words as data.” *American Political Science Review* 97(02):311–331.

Lazarsfeld, P.F., B. Berelson & H. Gaudet. 1948. “The People’s Choice. How the Voter Makes up His Mind in a Presidential Campaign [1944].”.

Lee Kaid, L, J Fernandes & D Painter. 2011. “Effects of Political Advertising in the 2008 Presidential Campaign.” *American Behavioral Scientist* 55(4):437–456.

Lodge, M., K.M. McGraw & P. Stroh. 1989. "An impression-driven model of candidate evaluation." *The American Political Science Review* pp. 399–419.

Lupia, A. & M.D. McCubbins. 2000. *Elements of reason: Cognition, choice, and the bounds of rationality.* Cambridge Univ Pr.

Madigan, D., A.E. Raftery, C. Volinsky & J. Hoeting. 1996. Bayesian model averaging. In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models, Portland, OR*. pp. 77–83.

Mann, T.E. & R.E. Wolfinger. 1980. "Candidates and parties in congressional elections." *The American political science review* pp. 617–632.

McCombs, M.E. & D.L. Shaw. 1993. "The evolution of agenda-setting research: Twenty-five years in the marketplace of ideas." *Journal of Communication* 43(2):58–67.

McCombs, M.E., D.L. Shaw & D.H. Weaver. 1997. *Communication and democracy: Exploring the intellectual frontiers in agenda-setting theory.* Lawrence Erlbaum.

Mendelberg, T. 1997. "Executing Hortons: Racial crime in the 1988 presidential campaign." *The Public Opinion Quarterly* 61(1):134–157.

Moffitt, R. 1993. "Identification and estimation of dynamic models with a time series of repeated cross-sections." *Journal of Econometrics* 59(1-2):99–123.

Monroe, B.L. & K. Maeda. 2004. Talk's cheap: Text-based estimation of rhetorical ideal-points. In *annual meeting of the Society for Political Methodology.* pp. 29–31.

Moon, W. 2006. "The paradox of less effective incumbent spending: Theory and tests." *British Journal of Political Science* 36(4):705.

Mutz, D.C. 1996. *Political persuasion and attitude change.* Univ of Michigan Pr.

Mutz, D.C. 1998. *Impersonal influence: How perceptions of mass collectives affect political attitudes.* Cambridge Univ Pr.

Mutz, D.C. 2007. "Effects of "in-your-face" television discourse on perceptions of a legitimate opposition." *American Political Science Review* 101(04):621–635.

Page, B.I. & R.Y. Shapiro. 1992. *The rational public: Fifty years of trends in Americans' policy preferences.* University of Chicago Press.

Page, S.E. 2008. *The difference: How the power of diversity creates better groups, firms, schools, and societies.* Princeton Univ Pr.

Pan, Z. & G.M. Kosicki. 1993. "Framing analysis: An approach to news discourse." *Political Communication* 10(1):55–75.

Petrocik, J.R. 1996. "Issue ownership in presidential elections, with a 1980 case study." *American Journal of Political Science* pp. 825–850.

Popkin, S.L. 1991. *The reasoning voter: Communication and persuasion in presidential campaigns.* University of Chicago Press.

Prior, M. 2007. *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections.* Cambridge Univ Pr.

Ridout, T.N., D.V. Shah, K.M. Goldstein & M.M. Franz. 2004. "Evaluating measures of campaign advertising exposure on political learning." *Political Behavior* 26(3):201–225.

Ridout, T.N. & M.M. Franz. 2011. *The persuasive power of campaign advertising.* Temple Univ Pr.

Riker, W.H. & Ordeshook, P.C. 1968. "A Theory of the Calculus of Voting." *The American political science review* 62(1):25–42.

Robson, A.R.W. 2005. "Multi-item contests." *Australian National University, Working Papers in Economics and Econometrics, WP* 446.

Schapire, R.E. 2003. "The boosting approach to machine learning: An overview." *Lecture Notes in Statistics* pp. 149–172.

Sebastiani, F. 2002. "Machine learning in automated text categorization." *ACM computing surveys (CSUR)* 34(1):1–47.

Segal, M.R. 2004. "Machine learning benchmarks and random forest regression.".

Shaw, D.R. 1999. "The effect of TV ads and candidate appearances on statewide presidential votes, 1988-96." *American Political Science Review* pp. 345–361.

Shaw, D.R. 2006. "The race to 270: The electoral college and the campaign strategies of 2000 and 2004.".

Sides, J. 2006. "The Origins of Campaign Agendas." *British Journal of Political Science* 36(03):407.

Sides, J. 2007. "The Consequences of Campaign Agendas." *American Politics Research* 35(4):465–488.

Sides, J. & A. Karch. 2008. "Messages that Mobilize? Issue Publics and the Content of Campaign Advertising." *The Journal of Politics* 70(02).

Sigelman, L. & E.H. Buell Jr. 2004. "Avoidance or engagement? Issue convergence in US presidential campaigns, 1960–2000." *American Journal of Political Science* 48(4):650–661.

Simon, A.F. 2002. *The winning message: Candidate behavior, campaign discourse, and democracy.* Cambridge Univ Pr.

Slapin, J.B. & S.O. Proksch. 2008. "A scaling model for estimating time-series party positions from texts." *American Journal of Political Science* 52(3):705–722.

Sniderman, P.M., R.A. Brody & P. Tetlock. 1993. *Reasoning and choice: Explorations in political psychology.* Cambridge Univ Pr.

Spiliotes, C.J. & L. Vavreck. 2002. "Campaign Advertising: Partisan Convergence or Divergence?" *Journal of Politics* 64(1):249–261.

Stevens, D. 2005. "Separate and Unequal Effects: Information, Political Sophistication and Negative Advertising in American Elections." *Political Research Quarterly* 58(3):413–425.

Stevens, D., J. Sullivan, B. Allen & D. Alger. 2008. "What's Good for the Goose is Bad for the Gander: Negative Political Advertising, Partisanship, and Turnout." *The Journal of Politics* 70(02).

Sulkin, T. & J. Evans. 2006. "Dynamics of diffusion." *American Politics Research* 34(4):505–534.

Tibshirani, R. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Vavreck, L. 2007. "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports." *Quarterly Journal of Political Science* 2(4):325–343.

Vavreck, L. 2009. *The message matters: the economy and presidential campaigns.* Princeton Univ Pr.

Westen, D. 2007. *The political brain: The role of emotion in deciding the fate of the nation.* Public Affairs.

Zaller, J. 1992. *The nature and origins of mass opinion.* Cambridge Univ Pr.

Table 1: Correlations between the ad coefficients assigned by the various text-based estimation methods.[a]

| | real | word reg | bayes | cut-off | KNN |
|---|---|---|---|---|---|
| word regression | 0.841* | | | | |
| bayes | 0.266* | 0.520* | | | |
| hard cut-off | -0.004 | 0.225* | 0.091 | | |
| KNN | 0.014 | 0.203* | 0.204* | 0.538* | |
| weighting | -0.006 | 0.277* | 0.198* | 0.940* | 0.681* |

[a] Values with * significant at $p < 0.05$.

Table 2: Top words for each CMAG category along with the success rate of the text-based classification technique.[a]

| (80% ads classified correctly) | | (86% ads classified correctly) | | (84% ads classified correctly) | |
|---|---|---|---|---|---|
| Attack | Promote | Personal | Policy | Democrat | Republican |
| jobs | country | war | jobs | jobs | voted |
| dollars | believe | country | tax | plan | taxes |
| tax | jobs | vietnam | plan | tax | war |
| iraq | people | believe | health | health | times |
| voted | healthcare | served | people | country | congress |
| national | plan | people | companies | american | people |
| taxes | american | truth | million | time | tax |
| war | health | life | iraq | companies | terrorists |
| million | world | time | insurance | healthcare | believe |
| care | family | world | care | iraq | health |

[a] The percentage is of those ads that belong to one of the two paired categories that were correctly assigned to the true CMAG category. For all three pairings, most ads had been put in one of these two categories; those that were not, were not included in the testing. Words are listed in descending absolute value of score.

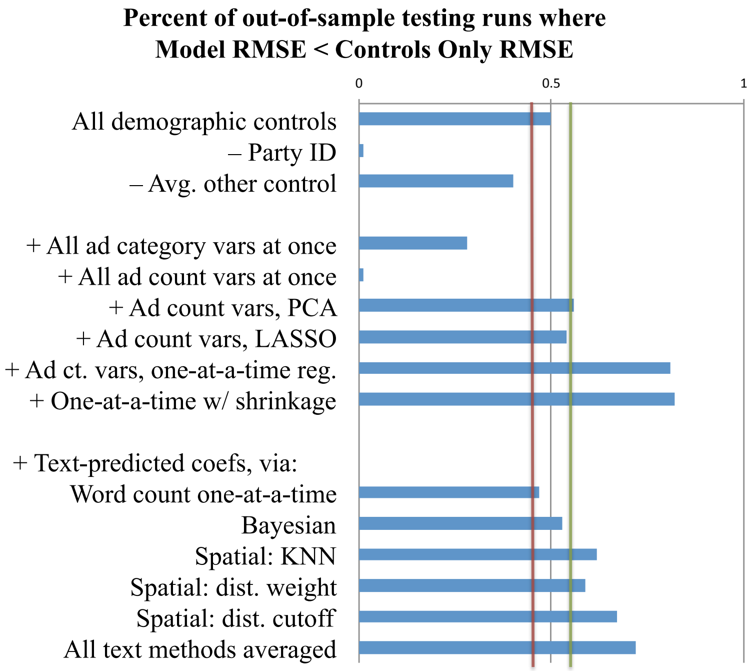

Figure 1: The length of each bar shows the proportion of runs in which the specified model out-performs the baseline controls-only model when comparing out-of-sample predictive accuracy. For each out-of-sample test, a model is fit on a random 95% of the observations, and then the dependent variable, vote intention, is predicted out-of-sample and compared with the actual measured vote intention. The baseline model predicts vote intention using only demographic variables (such as age, sex, party ID, etc). Each model in sections 2 and 3 adds on top of the demographic variables a method for measuring advertising effects. The more frequently a model beats the baseline, the better it is able to utilize the ad data, and the better it measures their effects. Longer bars are better; bars to the right of green band show predictions are significantly better than the controls-only baseline model at p $< 0.05$, and vice versa for bars to left of red.

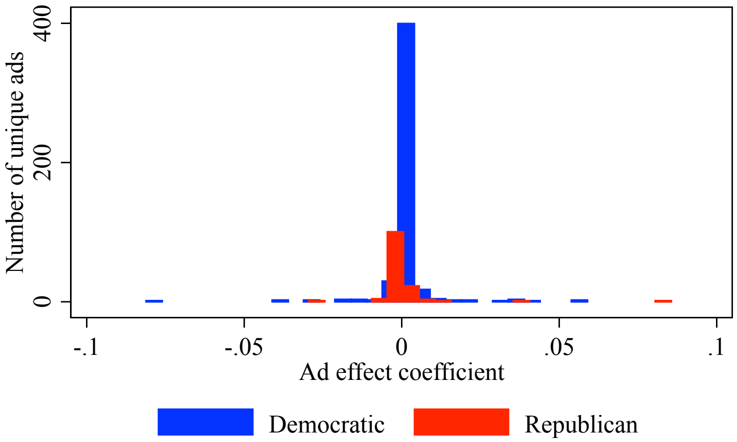

Figure 2: Distribution of ad effect coefficients, grouped by whether the ad was Democratic- or Republican-sponsored. Most ads have no effect, but many Republican-sponsored ads appear to have pro-Kerry effects, and vice versa.

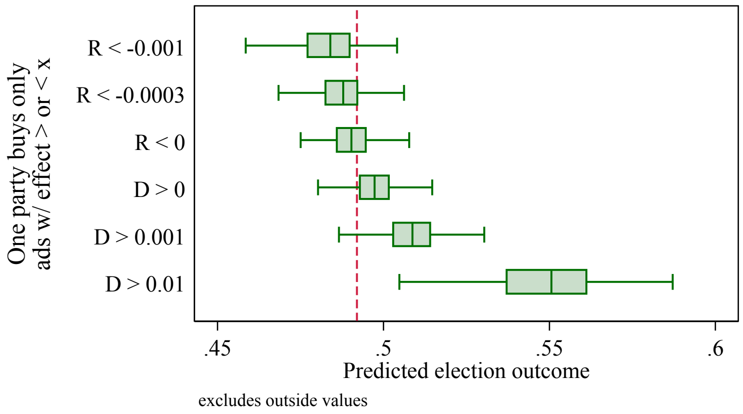

Figure 3: The predicted effect on (Democratic) vote share, if one party buys only the most effective (highest coefficient) ads. Dashed line is the actual outcome. Note that both parties do better if they only buy the most effective ads, but when using only their most effective ads, Democrats have much more room for improvement (bottom) than do Republicans (top). Errors are bootstrapped; shrunk coefficients are used.

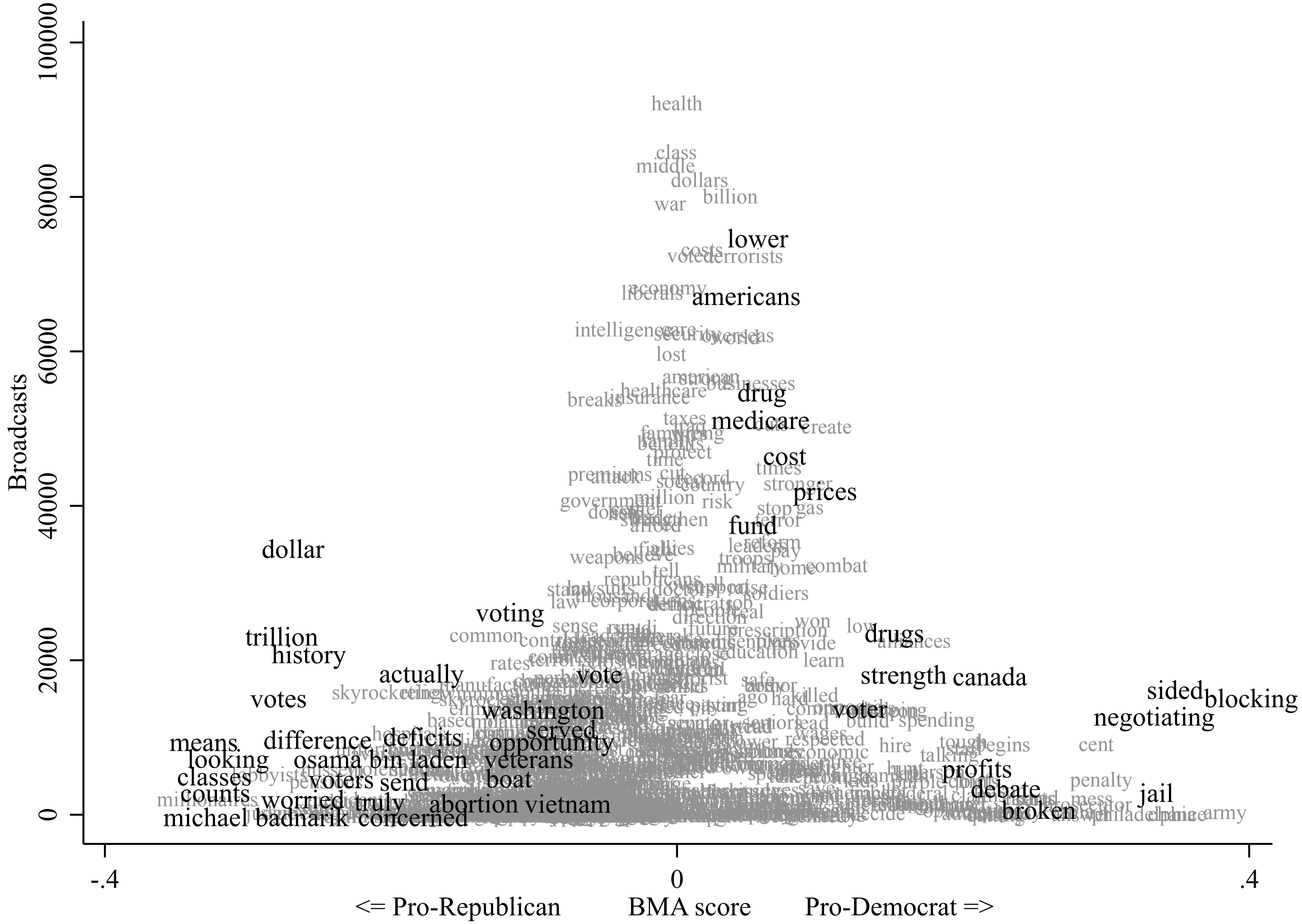

Figure 4: The top 1000 words plotted by each the word's BMA score and the number of broadcasts of ads containing it. Words in black are those with scores at least two standard deviations from 0. Note that positions have been jittered for legibility.

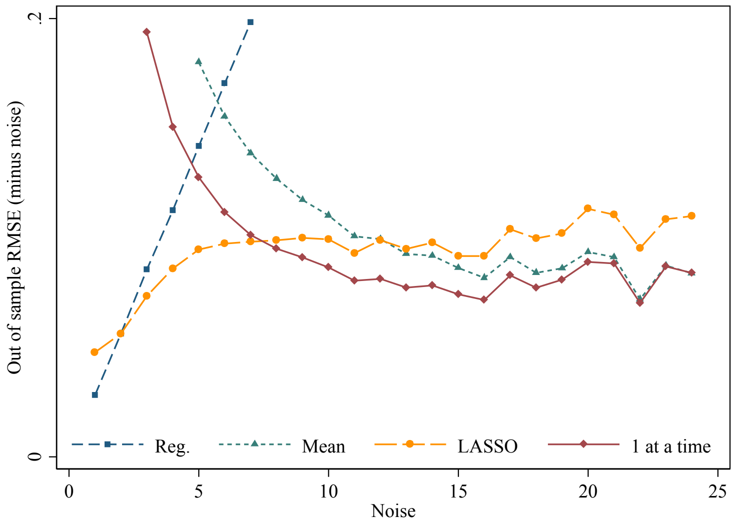

Figure 5: Monte Carlo simulations comparing the out-of-sample accuracy of four methods in estimating effects. For each simulation run, construct $y = \mathbf{X}\boldsymbol{\beta} + \epsilon$ with random X, $\beta$, and $\epsilon$ values, fit the specified model on a sub-sample, and predict $\hat{y}$ out of sample. Lower RMSE scores (y axis) indicate better predictive accuracy and thus more accurate estimation of the effects $\boldsymbol{\beta}$. Simulations are repeated 10,000 times for each level of noise $\epsilon \sim U[-m, m], m \in \{0....25\}$. Regression does best when noise is low, LASSO best when noise is medium, but one-at-a-time does best for a wide range of higher noise levels.

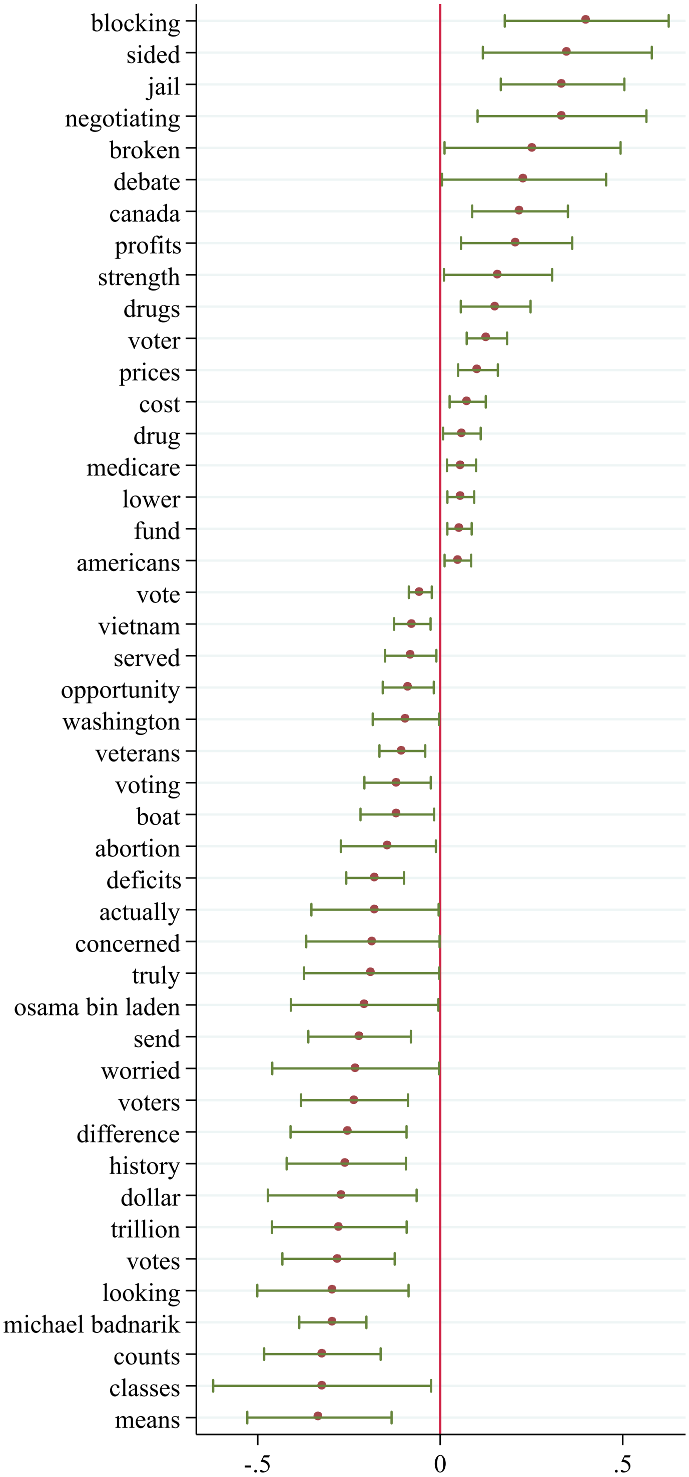

Figure 6: The top words with scores and 95% confidence intervals. Top = pro-Democratic.