Supplemental Materials

Deus ex Machina: Candidate Web Presence and the Presidential Nomination Campaign

Dino P. Christenson
Assistant Professor
Boston University
dinopc@bu.edu

Corwin D. Smidt
Assistant Professor
Michigan State University
smidtc@msu.edu

Costas Panagopoulos
Associate Professor
Fordham University
costas@post.harvard.edu

Contents

1 Measurement Appendix

1.1 Factors Included in the SIPP Index

The SIPP Index is a publicly-available measure developed by the Spartan Internet Consulting Corporation. The SIPP Index represents the aggregate score of a candidate's quantitative factors relative to overall market behavior. A broad-based set of 650 factors are used to provide an objective assessment of how each candidate is connecting with individuals through the Internet. These include the following:

- Candidate official website statistics, including: reach; page views per user; Google page rank; indexed pages; sites linked in; search engine optimization; Quantcast rank.
- Social media statistics: Myspace friends, comments, videos; Facebook supporters, wall posts, notes, groups; as well as activity on Eventful, Flickr, and Meetup.
- Youtube subscibers, views.
- Presence in online news outlets: Yahoo News, CNN, New York Times, Reuters, Fox News, Google News.
- Search engine ranking for key issues in Google, Yahoo, MSN, AOL, and Ask.com.
- Presence on blogs: Tecnorati number of blogs and number of blog posts.

In creating the index, each component score is weighted depending on its ability to connect to Internet users. For example, search engine rankings are rated less than activity on a candidate's own website. After tabulating and summing each candidate's activity across these measures, a relative score is made by dividing a candidate's total by the summed amount of activity for all candidates.

1.2 News Media and National Polling Data

Beginning in July of 2007 we employed daily automated content coding of newspaper websites.[1] We have collected national newspaper articles from the web using an automated retrieval program. Importantly, we only selected articles that were specifically located on each newspaper's "Campaign" webpage or RSS feed, such that we avoided selecting articles not specifically connected to the campaign.

Downloaded information includes the title, the journal, the time, and the content of the article, as well as the time of its posting. The program was written to retrieve information from a nonrandom sample of newspaper websites; collecting articles from all news sites across the country would indeed be worthwhile, but beyond realizable computational space. Given the necessity of a sample, we set out to balance the data collection across a few potential confounders, including timing of the primary, population size of the state, newspaper syndication size, and even the ideological tendencies of the newspapers. Here we focus on only our national sources, the *AP Politics Wire*, the *Washington Post*, the *New York Times*, and the *Los Angeles Times*, from the beginning of July 2007 through the last week of December 2007. This subset of data includes over 3,000 articles with more than 100,000 sentences spanning about 25 weeks.

After identifying all articles that covered the presidential nomination contest, we developed a measure of each candidate's news media prominence. Our measure counts the number of news media sentences that refer to any candidate for a party's nomination, and defines a candidate's media prominence as the percentage of sentences referring to that candidate out of all sentences discussing candidates of the corresponding party.

Our longitudinal media data is supplemented by a host of publicly available national polls. Prohibitive costs prevent polling by the same organization over the entire span of the

[1] Using LexisNexis other databases, we were able to ascertain a high level of similarity between a paper's print and online content. For evidence of the public's increasing use of online sources, and decreasing reliance on print sources within 2008 see the online Pew Report "Internet's Broader Role in Campaign 2008," January 11, 2008

campaign; rather it is common for various different organizations to purchase questions from a survey house to be administered at a particular time and state. Fortunately, a great deal of these polls ask a slight variant of the same question at different times, namely: "If the Republican/Democrat primary election were held today, who would you vote for - or who are you leaning toward today?" The consistent presence of this question in national polls during the primary provides the opportunity to create time series of electoral preferences and viability for each candidate.

Our models of the polls take into account differences in each poll's sample size and are capable of controlling for variants of question wording. In addition, when more than one poll is conducted at the same time we model the scores with a decrease in the standard error based on the combined sample size. Contrary to common practice, we are not concerned with who will win the primary or caucus, but how the electorate collectively changes their favored candidate(s) in covariance with news media coverage and other external events.

We include in our analysis any poll asking a close variant of the question: "If the Republican/Democrat primary election were held today, who would you vote for - or who are you leaning toward today?" The consistent presence of this question in state and national polls provides the opportunity to create time series of candidate standing from July 2007 through to the end of December. Our sample includes 187 national polls covering 80% and 84% of the days of our analysis. Our models of these polls take into account missing data and differences in each poll's sample size to weigh their relative accuracy. When more than one poll is in the field on the same day we take a weighted average of the two results based on sample sizes.

1.3 Weekly State Space Filtering

The use of daily measures of a candidate's polling level and media prominence suffer from both significant sampling error and frequent missing data. In the case of the

news media series we have few missing data but a small daily sample. In the case of the

public opinion series the opposite case is true; polls are not taken every day, but when they are taken they have relatively small levels of sampling error. To accommodate both these problems we specify a Bayesian state space model to estimate the underlying population parameters via Gibbs sampling. See Green, Gerber & Boef (1999) for further description of the intuition of these methods, in-depth treatments are found in Durbin & Koopman (2001) and West & Harrison (1997). Specifically, we follow Cargnoni, Müller & West (1997), and modify standard Bayesian linear state space model techniques to accomodate for compositional data of this type. For each model we use a random walk transition model with relatively diffuse priors.

References

Cargnoni, Claudia, Peter Müller & Mike West. 1997. "Bayesian Forecasting of Multinomial Time Series Through Conditionally Dynamic Models." *Journal of the American Statistical Association* 92:640–647.

Durbin, James & Sien Jan Koopman. 2001. *Time Series Analysis by State Space Methods.* Oxford: Oxford University Press.

Green, Donald, Alan Gerber & Suzanna De Boef. 1999. "Tracking Opinion Over Time: A Method for Reducing Sampling Error." *Public Opinion Quarterly* 63:178–192.

West, Mike & Jeff Harrison. 1997. *Bayesian Forecasting and Dynamics Models.* 2nd ed. New York: Springer.