Probabilistic Models in Political Science

Pablo Barberá
Center for Data Science
New York University
www.pablobarbera.com
George Takei
March 28 at 10:10pm ·

Who's with me.

Bon Alimagno
@karma_thief

I need a hug. I have never been so traumatized by a television show.
#gameofthrones

10:06 PM - 2 Jun 2013

Sophie

Last night I got so drunk I got kicked out of a club within an hour, cried and called my parents and got them to pick me up at 2am. Hockey would be proud.

Like · Share

6 people like this.

Jenny

I am very proud of you!

8 hrs · Like

Justin Bieber
@justinbieber

I make music. I love music.

10:09 PM - 7 Apr 2014
The harmonious development of Crimea and Sevastopol as part of our state is one of the main objectives of the Russian Government.

10:39 AM - 21 Mar 2014

Much of the foreign media coverage has distorted the reality of my country and the facts surrounding the events," writes Nicolás Maduro, the president of Venezuela, in Opinion: http://nyti.ms/1gP5o2l

Like · Comment · Share

262 people like this.

I'm not giving up on our fight to extend unemployment benefits. Watch my interview with Now With Alex Wagner about why we need to keep fighting.

1:54 PM - 16 Jan 2017

Warren: This is the moment to back on economy
www.msnbc.com
President Obama faces one huge problem with his effort to improve the economy: an opposition party

Like · Comment · Share

15,483 720 1,041

Today, a representative from my office will be meeting with constituents in Goshen. For more details, visit walorski.house.gov/services/upcom...

11:22 AM - 8 Apr 2014

Like · Comment · Share

57
Two approaches to the study of social media and politics:

1. How social media platforms transform political communication
   - Are social media creating ideological “echo chambers”?

2. Social media as digital traces of political behavior
   - Can we infer latent individual traits (e.g. political ideology) from online ties (follows, likes...)?
Inferring political ideology using Twitter data

- Two common patterns about social behavior:
  1. Homophily: clustering in social networks along common traits (“birds of a feather tweet together”)
  2. Selective exposure: preference for information that reinforces current views and for avoiding opinion challenges.

- Social media networks replicate offline networks.

- **Key assumption:** individuals prefer to follow political accounts they perceive to be ideologically close.

- These decisions contain information about allocation of scarce resource (attention).

- Use this information to estimate ideological locations of politicians *and* individuals on the latent same scale.
Political Accounts

- FiveThirtyEight
- WhiteHouse
- BarackObama
- NYTimeskrugman
- HRC
- senrobportman
- maddow

| pol. account | BarackObama | WhiteHouse | GOP | maddow | FoxNews | HRC | ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ryanpetrik</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>user 2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>user 3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>user 4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>user 5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>user n</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>
Spatial following model

- Users’ and politicians’ ideology ($\theta_i$ and $\phi_j$) are defined as latent variables to be estimated.
- Data: “following” decisions, a matrix of binary choices ($Y_{ij}$).
- Spatial following model: for $n$ users, indexed by $i$, and $m$ political accounts, indexed by $j$:

$$P(y_{ij} = 1 | \alpha_j, \beta_i, \gamma, \theta_i, \phi_j) = \text{logit}^{-1} \left( \alpha_j + \beta_i - \gamma(\theta_i - \phi_j)^2 \right)$$

where:

- $\alpha_j$ measures *popularity* of politician $j$
- $\beta_i$ measures *political interest* of user $i$
- $\gamma$ is a normalizing constant
Intuition of the model

Probability that Twitter user $i$ follows politician $j$, as a function of the user’s ideology:

$$\Pr(y_{ij} = 1)$$

- $\phi_1 = -1.51$, $\alpha_1 = 3.51$
- $\phi_2 = 1.09$, $\alpha_2 = 2.59$
Estimation

- Goal of learning:
  - $\theta_i$: ideological positions of users $i = 1, \ldots, n$
  - $\phi_j$: ideological positions of political accounts $j = 1, \ldots, m$

- Likelihood function:

$$p(y|\theta, \phi, \alpha, \beta, \gamma) = \prod_{i=1}^n \prod_{j=1}^m \logit^{-1}(\pi_{ij})^{y_{ij}}(1 - \logit^{-1}(\pi_{ij}))^{1-y_{ij}}$$

where $\pi_{ij} = \alpha_j + \beta_i - \gamma(\theta_i - \phi_j)^2$

- Exact inference is intractable $\rightarrow$ MCMC (approx. inference)

- Estimation:
  - First stage: HMC in Stan with random sample of $Y$ to compute posterior distribution of $j$-indexed parameters.
  - Second stage: parallelized MH in R for rest of $i$-indexed parameters (assuming independence), on NYU’s HPC.
Data

- $m =$ list of 620 popular political accounts in the U.S.
  - Legislators, president, candidates, other political figures, media outlets, journalists, interest groups...

- $n =$ followers of at least one of these accounts
  - 30.8M users (~75% of U.S. users)
  - 100K of these were matched with voter files
    - States: AK, CA, FL, OH, PA.
    - Unique, perfect matches on first and last name, and county.

- Code:
  - Method: github.com/pablobarbera/twitter_ideology
  - Applications: github.com/SMAPPNYU/echo_chambers
  - Data collection: streamR, Rfacebook packages for R (available on CRAN)
  - Data analysis: github.com/pablobarbera/pytwools (python)
Results

Political Actors
@sentedcruz
Median House R
Median Senate R
@senjohnmccain
Median Senate D
Median House D
@BarackObama
@VP
@nancypelosi
@HillaryClinton
@sensanders

Media
@limbaugh
@glennbeck
@DRUDGE_REPORT
@FoxNews
@washingtonpost
@cnnbrk
@nytimes
@msnbc
@NPR
@maddow
@motherjones

Interest Groups
@redstate
@nra
@Heritage
@AEI
@CatoInstitute
@RANDCorporation
@BrookingsInst
@hrw
@aclu
@dailykos
@ OccupyWallSt
@glaad
@HRC

Position on latent ideological scale
Validation

This method is able to correctly classify and scale Twitter users on the left-right dimension:

1. Political accounts
   - Correlation with measures based on roll-call votes.

2. Ordinary citizens
   - Individual and aggregate-level survey responses
   - Voting registration files

It is also able to predict change over time.
Ideal Points of Members of the 113th U.S. Congress

Political elites

Estimated Twitter Ideal Points

House

Senate

\[ \rho_D = 0.63 \]

\[ \rho_R = 0.46 \]

\[ \rho_D = 0.66 \]

\[ \rho_R = 0.63 \]
Ordinary Users

Comparison with ideology estimates from aggregated surveys (Lax and Phillips, 2012; Tausanovitch and Warshaw, 2013)

Mean Liberal Opinion (Lax and Phillips, 2012)

Public Preference Estimate (Tausanovitch and Warshaw, 2013)

\[ \rho = -0.916 \]

\[ \rho = 0.791 \]
Ordinary Users

Republicans are more conservative than Democrats

Predictive accuracy for party affiliation is 83%
Barberá “Who is the most conservative Republican candidate for president?” *The Washington Post*, June 16 2015
Application: Twitter as an Ideological Echo Chamber?

Barberá (2015) “Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data.” Political Analysis
Application: Twitter as an Ideological Echo Chamber?

Other applications

Ideology of media outlets
Ideological Asymmetries
Multidimensional Policy Spaces
Two approaches to the study of social media and politics:

1. How social media platforms transform political communication
   - As voters are able to directly interact with politicians, does the quality of political representation improve?

2. Social media as digital traces of political behavior
   - Are legislators’ and citizens’ social media messages a valid proxy for the attention they give to different political issues?
Political Representation

Public Opinion → Policy
Political Representation

Issues Voters Discuss → Issues Legislators Discuss

Do Legislators Accurately Represent Voters’ Interests?
Who Leads? Who Follows?

Outline

1. Analyze tweets sent by Members of U.S. Congress and their followers using topic modeling techniques.
2. Estimate the importance (frequency of discussion) of 100 different issues in the revealed expressed political agenda for legislators and constituents
3. **Political Congruence:** are Members of Congress discussing the same set of issues as their constituents?
4. **Political Responsiveness:** do topics discussed by Members of Congress temporally precede or follow topics discussed by the voters?
651,116 tweets by Members of U.S. Congress, from Jan. 1, 2013 to Dec. 31, 2014 (113th Congress), collected by the Social Media and Political Participation Lab (SMaPP) using Twitter’s Streaming API.
Citizens’ Tweets

Collected all tweets for 3 samples of citizens:

1. Informed public:
   - Followers of 5 major media outlets (CNN, FoxNews, MSNBC, NYT, WSJ) located in U.S. (filtered by time zone)
   - Random sample of 10,000 (out of ~30M)

2. Republican Party Supporters:
   - Follow 3+ Rep MCs and no Dem MCs
   - Random sample of 10,000 (out of 203,140)

3. Democratic Party Supporters:
   - Follow 3+ Dem MCs and no Rep MCs
   - Random sample of 10,000 (out of 67,843)
**Table: Number of tweets in dataset**

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Avg</th>
<th>Min</th>
<th>Max</th>
<th>Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>House Republicans</td>
<td>238</td>
<td>1,215</td>
<td>70</td>
<td>8,857</td>
<td>267,311</td>
</tr>
<tr>
<td>House Democrats</td>
<td>207</td>
<td>1,177</td>
<td>113</td>
<td>5,993</td>
<td>222,491</td>
</tr>
<tr>
<td>Senate Republicans</td>
<td>46</td>
<td>1,532</td>
<td>73</td>
<td>6,627</td>
<td>67,412</td>
</tr>
<tr>
<td>Senate Democrats</td>
<td>56</td>
<td>1,616</td>
<td>150</td>
<td>10,736</td>
<td>87,307</td>
</tr>
<tr>
<td>Informed Public</td>
<td>10K</td>
<td>948</td>
<td>2</td>
<td>5,861</td>
<td>9,487,382</td>
</tr>
<tr>
<td>Rep. Supporters</td>
<td>10K</td>
<td>1,091</td>
<td>2</td>
<td>8,804</td>
<td>10,911,813</td>
</tr>
<tr>
<td>Dem. Supporters</td>
<td>10K</td>
<td>1,306</td>
<td>2</td>
<td>5,122</td>
<td>13,058,947</td>
</tr>
</tbody>
</table>

Period of analysis: January 1, 2013 to December 31, 2014.
Political Representation

Media data:
- 273,007 tweets from 36 largest media outlets in U.S. (print, broadcast, online) over same period.
From Tweets to Topics

4 steps in our analysis

1. Tweets from Members of Congress are preprocessed and split by day, party and chamber (N=2,920 documents)

2. Latent Dirichlet Allocation (Blei, 2003):
   - Each document is a mixture over $K = 100$ latent topics.
   - Topics are distributions over $V = 75,000$ n-grams (up to trigrams, selected by frequency; keeping hashtags)
   - Estimated parameters:
     - $\hat{\beta}$ Distribution of n-grams over topics ($K \times V$)
     - $\hat{\theta}$ Distribution of topics over documents ($K \times N$)

3. Similar text processing for tweets from citizens and NYT tweets (split by day and group)

4. Using simulation, compute posterior distribution of $\hat{\theta}_F$ for observed n-grams for citizens and media
Latent Dirichlet allocation (LDA)

- **Topic models** are powerful tools for exploring large data sets and for making inferences about the content of documents.

- Many applications in information retrieval, document summarization, and classification.

- LDA is one of the simplest and most widely used topic models.
Latent Dirichlet Allocation

- Document = random mixture over latent topics
- Topic = distribution over n-grams

Probabilistic model with 3 steps:
1. Choose $\theta_i \sim \text{Dirichlet}(\alpha)$
2. Choose $\beta_k \sim \text{Dirichlet}(\delta)$
3. For each word in document $i$:
   - Choose a topic $z_m \sim \text{Multinomial}(\theta_i)$
   - Choose a word $w_{im} \sim \text{Multinomial}(\beta_{i,k=z_m})$

where:
- $\alpha$=parameter of Dirichlet prior on distribution of topics over docs.
- $\theta_i$=topic distribution for document $i$
- $\delta$=parameter of Dirichlet prior on distribution of words over topics
- $\beta_k$=word distribution for topic $k$
Applications that aggregate by author or day outperform tweet-level analyses (Hong and Davidson, 2010).

K is fixed at 100 based on cross-validated model fit.

Text is parsed with scikit-learn in python.

Estimation: Collapsed Gibbs Sampler in C++ (Griffits and Steyvers, 2004), ported to \( \mathbb{R} \) by Grün and Hornik (2011)
Validation

j.mp/lda-congress-demo
Congruence

Are Members of Congress discussing the same set of issues as their constituents?

Table: Contemporaneous Pearson Correlations in Topic Distribution

<table>
<thead>
<tr>
<th>Group</th>
<th>Dem Mcs</th>
<th>Rep MCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democratic Members of Congress</td>
<td>1.00</td>
<td>0.22</td>
</tr>
<tr>
<td>Republican Members of Congress</td>
<td>0.22</td>
<td>1.00</td>
</tr>
<tr>
<td>Informed Public</td>
<td>0.33</td>
<td>0.39</td>
</tr>
<tr>
<td>Republican Party Supporters</td>
<td>0.17</td>
<td>0.62</td>
</tr>
<tr>
<td>Democratic Party Supporters</td>
<td>0.58</td>
<td>0.33</td>
</tr>
<tr>
<td>Media</td>
<td>0.39</td>
<td>0.61</td>
</tr>
</tbody>
</table>
Do legislators influence the public? Does the public influence legislators?

To explore causal relationships between topic distributions, we use a Granger-causality framework (Granger, 1969):

- Regress proportion of tweets on topic $k$ at time $t$ by each group on lagged proportions for all groups, using five lags.
- Do legislators’ tweets predict tweets by the public, controlling for the media, and vice versa?
- Changes in tweets as proxies for changes in salience of issues
Results: Democratic legislators

Diagram:

- Public
- Media
- Legislators

Arrows and Values:
- Public to Legislators: <0.00
- Public to Media: 0.25
- Media to Public: 0.01
- Legislators to Media: 0.09
- Legislators to Public: 0.04
Results: Republican legislators
Conclusions

1. Social media as variable
2. Social media as data

Future work / open questions:

- More complex generative models for tweets that exploit platform features (Author-Topic; Dynamic; Hierarchical)
- Text- vs network-based estimates of political ideology
- Predicting latent probability to turn out to vote based on tweet text, using voting registration records
- Multilingual topic modeling
- Detecting irony and sarcasm (Trump!)
- Identifying bots and spam with user and text features only
Thanks!

website: pablobarbera.com
twitter: @p_barbera
github: pablobarbera
Model with covariates
Model identification
Unequal representation
Comparative responsiveness
Model with Covariates

Baseline model:

\[ P(y_{ij} = 1) = \text{logit}^{-1} \left( \alpha_j + \beta_i - \gamma (\theta_i - \phi_j)^2 \right) \]

Model with geographic covariate:

\[ P(y_{ij} = 1) = \text{logit}^{-1} \left( \alpha_j + \beta_i - \gamma (\theta_i - \phi_j)^2 + \delta s_{ij} \right) \]

where \( s_{ij} = 1 \) if user \( i \) and political actor \( j \) are located in the same state, and \( s_{ij} = 0 \) otherwise.

\[ \hat{\delta} \approx 1.20 \text{ and } \hat{\gamma} \approx 0.90 \]
Comparing Parameter Estimates Across Different Model Specifications

<table>
<thead>
<tr>
<th>$\phi_i$, Elites' Ideology Estimates</th>
<th>$\theta_i$, Users' Ideology Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter Estimates (Baseline Model)</td>
<td>Parameter Estimates (Model with Geographic Covariate)</td>
</tr>
</tbody>
</table>

-2  -1  0  1  2

-2  -1  0  1  2
Identification

\[ P(y_{ijt} = 1) = \logit^{-1} \left( \alpha_j + \beta_i - \gamma(\theta_{it} - \phi_j)^2 \right) \]

Additive aliasing:

\[ = \logit^{-1} \left( (\alpha_j + k) + (\beta_i - k) - \gamma(\theta_{it} - \phi_j)^2 \right) \]

\[ = \logit^{-1} \left( \alpha_j + \beta_i - \gamma((\theta_{it} + k) - (\phi_j - k))^2 \right) \]

Multiplicative aliasing:

\[ = \logit^{-1} \left( \alpha_j + \beta_i - \frac{\gamma}{k^2}((\theta_{it} - \phi_j) \times k)^2 \right) \]
## Identifying restrictions

<table>
<thead>
<tr>
<th>Indeterminacy</th>
<th>Approach 1</th>
<th>Approach 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive aliasing (1)</td>
<td>Fix $\alpha'_i = 0$ or $\beta'_i = 0$</td>
<td>Fix $\mu_\alpha = 0$ or $\mu_\beta = 0$</td>
</tr>
<tr>
<td>Additive aliasing (2)</td>
<td>Fix $\phi'_i = +1$ or $\theta'_i = +1$</td>
<td>Fix $\mu_\phi = 0$ or $\mu_\theta = 0$</td>
</tr>
<tr>
<td>Multiplicative aliasing</td>
<td>Fix $\phi''_i = -1$ or $\theta''_i = -1$</td>
<td>Fix $\sigma_\phi = 1$ or $\sigma_\theta = 1$</td>
</tr>
</tbody>
</table>

![Graph showing the relationship between true and estimated values of phi and theta for both approaches](image-url)

- **Approach 1**: Fix $\alpha'_i = 0$ or $\beta'_i = 0$ to resolve additive aliasing (1).
- **Approach 2**: Fix $\mu_\alpha = 0$ or $\mu_\beta = 0$ to resolve additive aliasing (1).
- **Approach 1**: Fix $\phi'_i = +1$ or $\theta'_i = +1$ to resolve additive aliasing (2).
- **Approach 2**: Fix $\mu_\phi = 0$ or $\mu_\theta = 0$ to resolve additive aliasing (2).
- **Approach 1**: Fix $\phi''_i = -1$ or $\theta''_i = -1$ to resolve multiplicative aliasing.
- **Approach 2**: Fix $\sigma_\phi = 1$ or $\sigma_\theta = 1$ to resolve multiplicative aliasing.
Barberá & Sood (2014) “Follow Your Ideology: A Measure of Ideological Location of Media Sources”, *MPSA Conference*
Barberá & Sood (2014) “Follow Your Ideology: A Measure of Ideological Location of Media Sources”, *MPSA Conference*
Application: Multidimensional Policy Spaces in Europe

\[ P(y_{ij} = 1) = \logit^{-1} \left( \alpha_i + \beta_j - \sum_{k=1}^{d} \gamma_d (\theta_{ik} - \phi_{jk})^2 \right) \]

Estimated ideological positions for 120 parties in 28 European countries

Unequal representation

We also analyze whether correspondence between citizens and legislators is higher for:

- Co-partisans (party supporters)
- Issues *owned* by each party (e.g. economy for Republicans; social issues for Democrats)
- Constituents (vs general public)
- Informed public vs random sample of U.S. Twitter users
- Individuals with income above median
What institutional configurations foster better representation?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany</td>
<td>Coalition</td>
<td>Prop.</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Spain</td>
<td>Single-party</td>
<td>Prop.</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>UK</td>
<td>Coalition</td>
<td>Maj.</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>France</td>
<td>Single-party</td>
<td>Maj.</td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

Electoral Institutions and Political Representation

j.mp/EPSA-lda-demo