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.

Like · Comment · Share

408,735 people like this.

66,990 shares

Bon Alimagno
@karma_thief

Follow

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

Reply Retweet Favorite ··· More

RETWEETS 356 FAVORITES 110

10:06 PM - 2 Jun 2013

Sophie ______
8 hrs

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. ✓ Seen by 24

Jenny ____ I am very proud of you!
8 hrs · Like

Justin Bieber
@justinbieber

Follow

I make music. I love music.

Reply Retweet Favorite ··· More

RETWEETS 54,213 FAVORITES 59,205

10:09 PM - 7 Apr 2014

Dmitry Medvedev @MedvedevRussiaE
Follow

The harmonious development of Crimea and Sevastopol as part of our state is one of the main objectives of the Russian Government

Reply · Retweet · Favorite · More

RETWEETS 144 FAVORITES 57

10:39 AM - 21 Mar 2014

The New York Times
April 2

"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/1gP5o2I

Like · Comment · Share 57

262 people like this. Top Comments

Elizabeth Warren shared a link.
January 16

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.

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

Jackie Walorski @RepWalorski
Follow

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

Reply · Retweet · Favorite · More

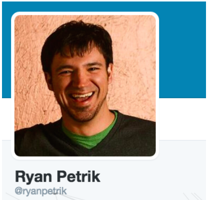
11:22 AM - 8 Apr 2014

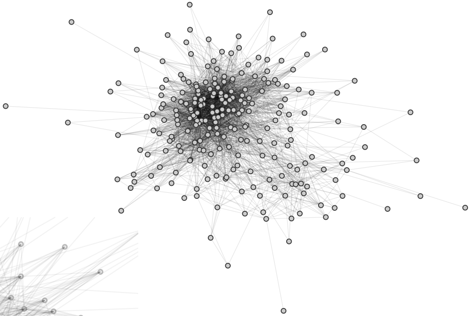
Two approaches to the study of social media and politics:

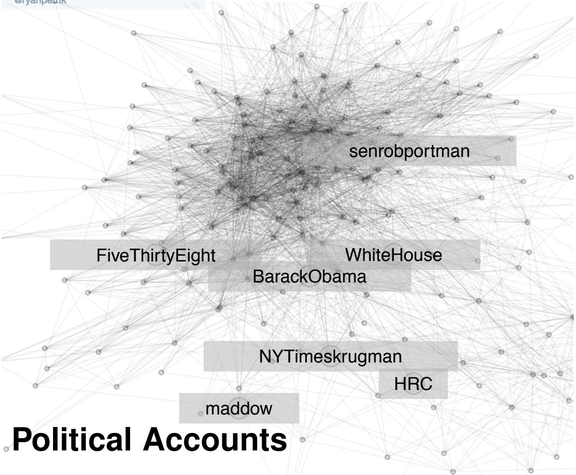
1. How social media platforms transform political communication
   - ▷ **Are social media creating ideological "echo chambers"?**
2. Social media as digital traces of political behavior
   - ▷ **Can we infer latent individual traits (e.g. political ideology) from online ties (follows, likes...)?**

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

	BarackObama	WhiteHouse	GOP	maddow	FoxNews	HRC	...	pol. account m
ryanpetrik	1	1	0	1	0	1	...	
user 2	0	0	1	0	1	0	...	
user 3	0	0	1	0	1	0	...	
user 4	1	1	0	0	0	1	...	
user 5	0	1	0	0	0	1	...	
...								
user n	0	1	1	0	0	0	...	

Spatial following model

- Users' and politicians' ideology (θ_i and ϕ_j) are defined as latent variables to be estimated.
- Data: "following" decisions, a matrix of binary choices ($\mathbf{Y_{ij}}$).
- 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:

α_j measures *popularity* of politician *j*

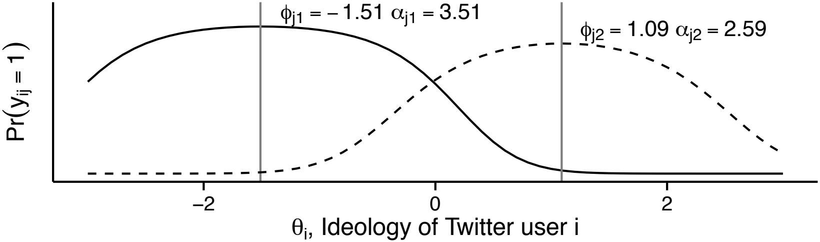
β_i measures *political interest* of user *i*

γ is a normalizing constant

More

Intuition of the model

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



Estimation

- Goal of learning:
  - θ_i: ideological positions of users $i = 1, \ldots, n$
  - ϕ_j: ideological positions of political accounts $j = 1, \ldots, m$
- Likelihood function:

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

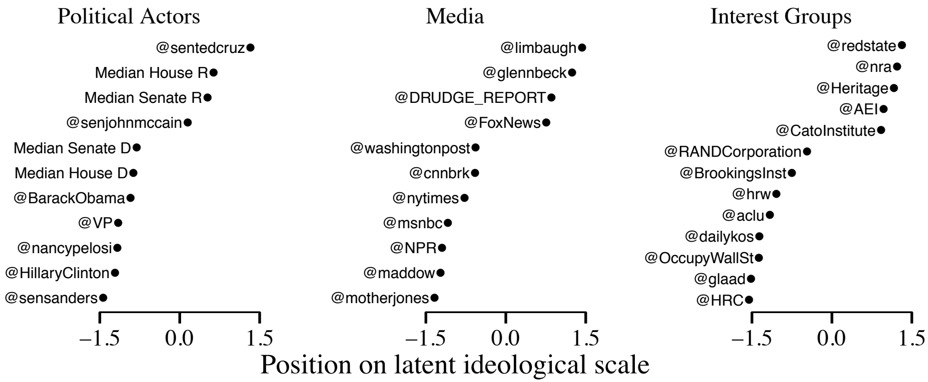
$$\text{where } \pi_{ij} = \alpha_j + \beta_i - \gamma(\theta_i - \phi_j)^2$$

- 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.
  - $\rightarrow$ Legislators, president, candidates, other political figures, media outlets, journalists, interest groups. . .
- n = followers of at least one of these accounts
  - $\rightarrow$ 30.8M users ($\sim$75% of U.S. users)
  - $\rightarrow$ 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
–1.5
0.0
1.5
Position on latent ideological scale

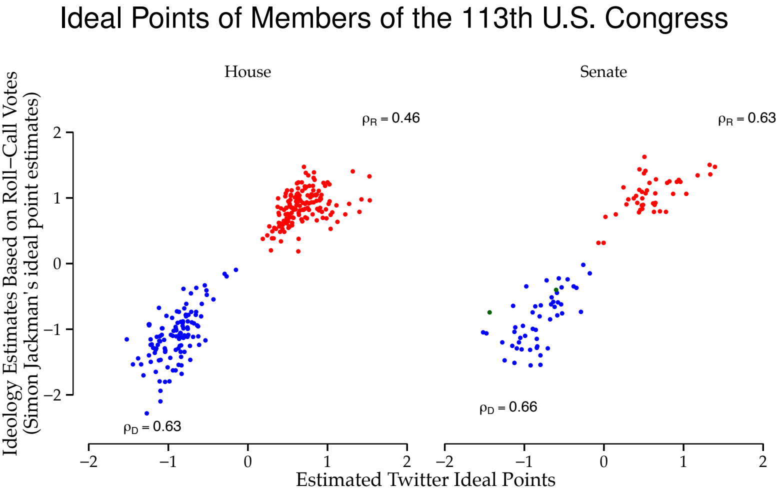
Validation

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

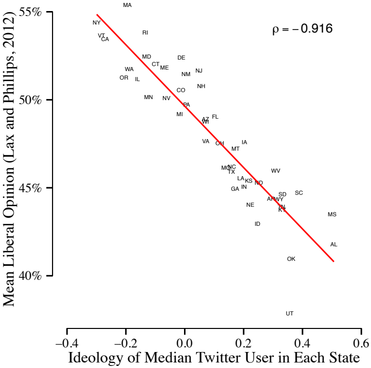
It is also able to predict change over time.

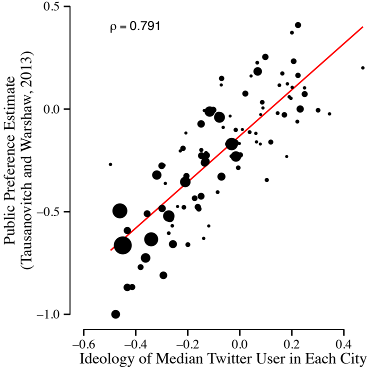
Political elites



Ordinary Users

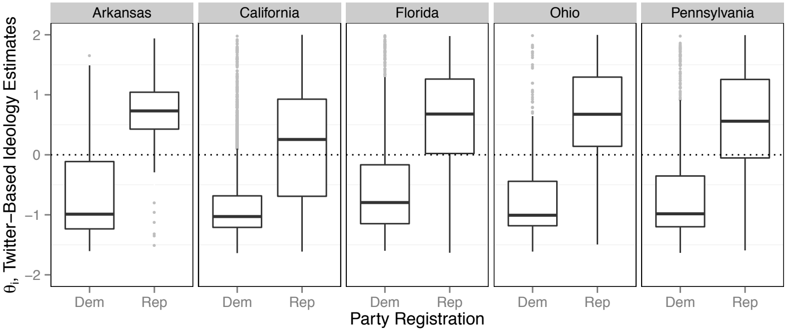
Comparison with ideology estimates from aggregated surveys (Lax and Phillips, 2012; Tausanovitch and Warshaw, 2013)





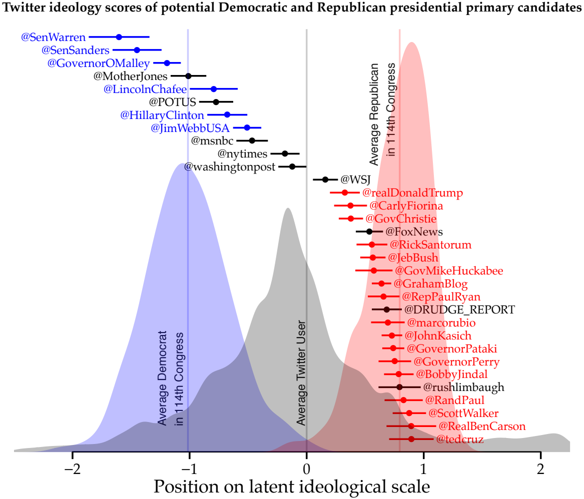
Ordinary Users

Republicans are more conservative than Democrats



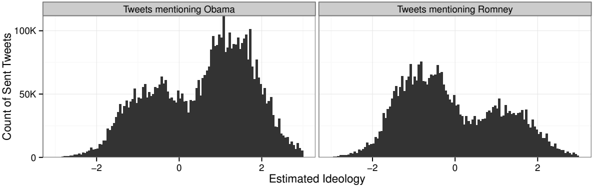
Predictive accuracy for party affiliation is 83%

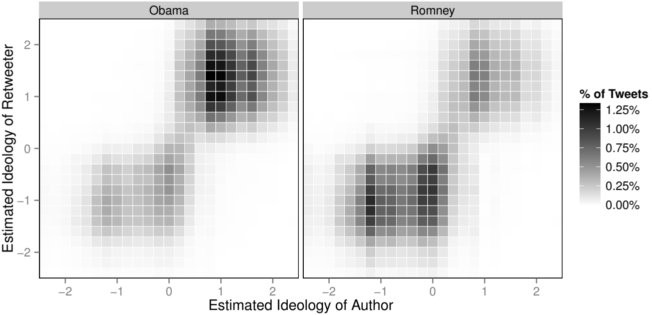
Application: Ideology of Presidential Candidates



Barberá "Who is the most conservative Republican candidate for president?" *The Washington Post*, June 16 2015

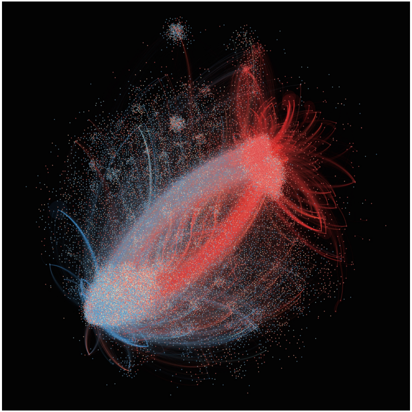
Application: Twitter as an Ideological Echo Chamber?

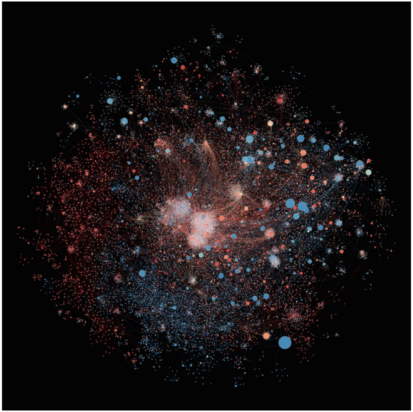




Barberá (2015) "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis*

Application: Twitter as an Ideological Echo Chamber?





Barberá, Jost, Nagler, Tucker, & Bonneau (2015) “Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?” *Psychological Science*

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?**

   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 $\longrightarrow$ Policy

Political Representation

$$\begin{tikzcd} \text{Issues Voters Discuss} \arrow[r] & \text{Issues Legislators Discuss} \arrow[l, bend left] \end{tikzcd}$$

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

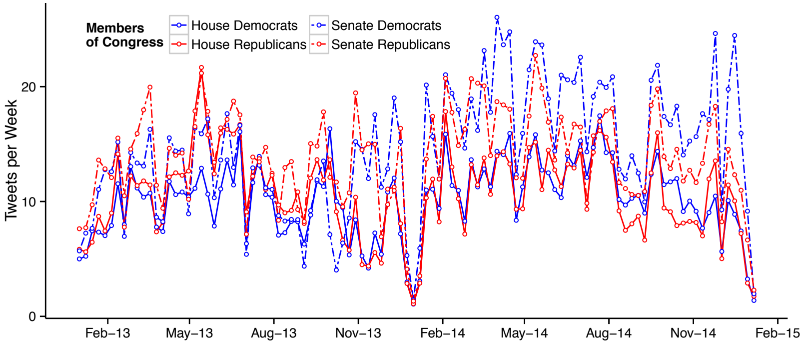
Barberá, Nagler, Egan, Bonneau, Jost, & Tucker (2014) "Leaders or Followers? Measuring Political Responsiveness in the U.S. Congress Using Social Media Data." APSA Conference Paper.

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?

Data

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:

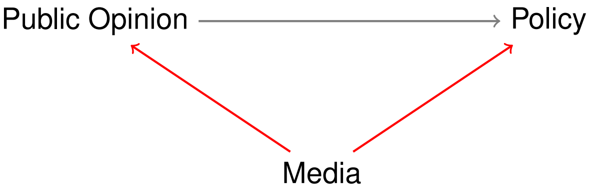
1. Informed public:
   - Followers of 5 major media outlets (CNN, FoxNews, MSNBC, NYT, WSJ) located in U.S. (filtered by time zone)
   - Random sample of 10,000 (out of $\sim$30M)
2. Republican Party Supporters:
   - Follow 3+ Rep MCs and no Dem MCs
   - Random sample of 10,000 (out of 203,140)
3. Democratic 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

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

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

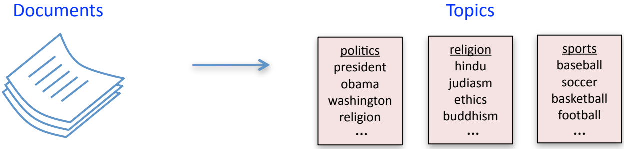
1. Tweets from Members of Congress are preprocessed and split by day, party and chamber (N=2,920 documents)
2. Latent Dirichlet Allocation (Blei, 2003):
   - Each document is a mixture over $K = 100$ latent topics.
   - Topics are distributions over $V = 75,000$ n-grams (up to trigrams, selected by frequency; keeping hashtags)
   - Estimated parameters:

     $\hat{\beta}$ Distribution of n-grams over topics ($K \times V$)

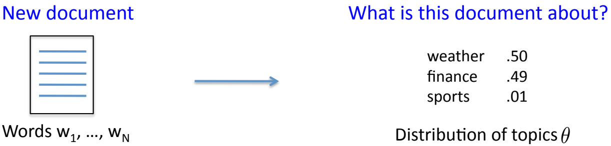
     $\hat{\theta}$ Distribution of topics over documents ($K \times N$)
3. Similar text processing for tweets from citizens and NYT tweets (split by day and group)
4. Using simulation, compute posterior distribution of $\hat{\theta}_F$ for observed n-grams for citizens and media

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

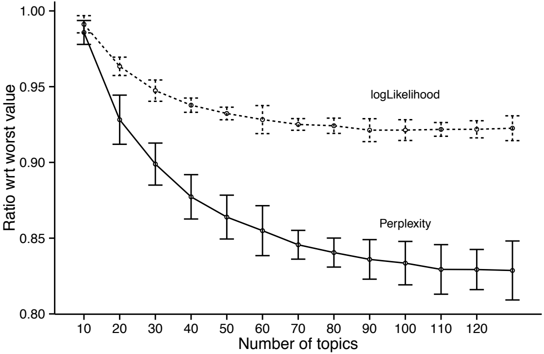
Probabilistic model with 3 steps:

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

where:
α=parameter of Dirichlet prior on distribution of topics over docs.
θ_i=topic distribution for document *i*
δ=parameter of Dirichlet prior on distribution of words over topics
β_k=word distribution for topic *k*

Estimation

- 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 `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

Group	Dem Mcs	Rep MCs
Democratic Members of Congress	1.00	0.22
Republican Members of Congress	0.22	1.00
Informed Public	0.33	0.39
Republican Party Supporters	0.17	0.62
Democratic Party Supporters	0.58	0.33
Media	0.39	0.61

Responsiveness

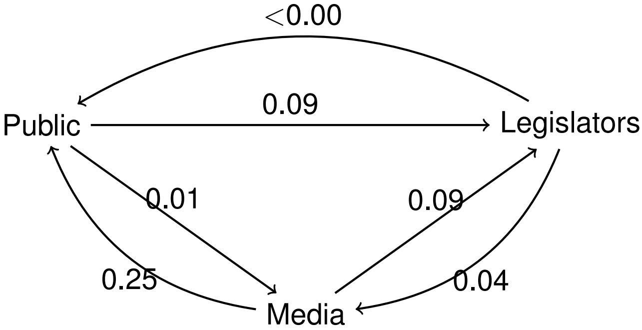
Do legislators influence the public? Does the public influence legislators?

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

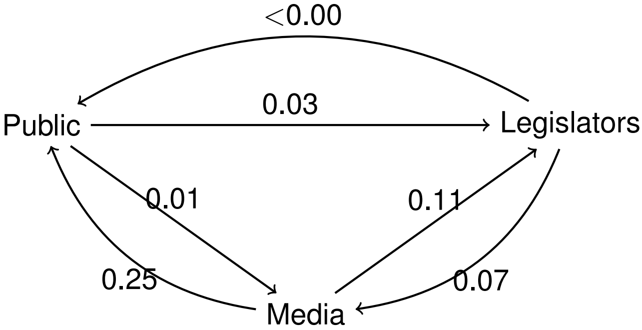
- Regress proportion of tweets on topic k at time t by each group on lagged proportions for all groups, using five lags.
- Do legislators' tweets predict tweets by the public, controlling for the media, and vice versa?

$\rightarrow$ Changes in tweets as proxies for changes in salience of issues

Results: Democratic legislators



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

Backup slides (index)

Model with Covariates

Baseline model:

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

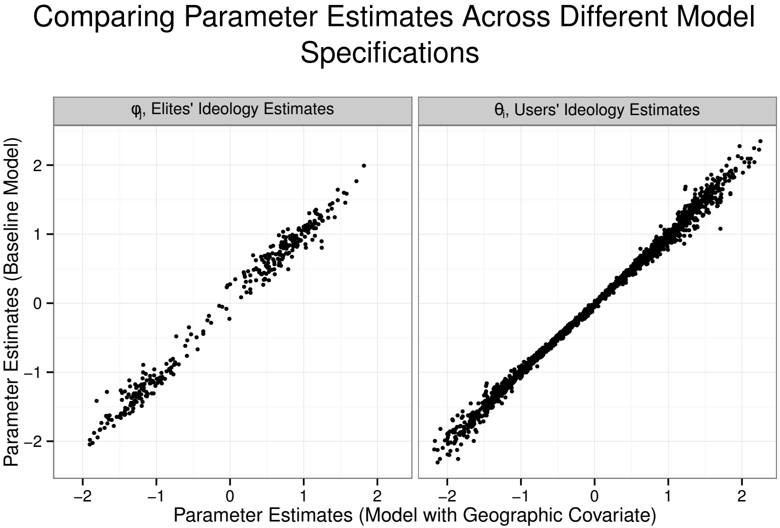
Model with geographic covariate:

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

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

$\hat{\delta} \approx 1.20$ and $\hat{\gamma} \approx 0.90$

Model with Covariates



Index

Identification

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

Additive aliasing:

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

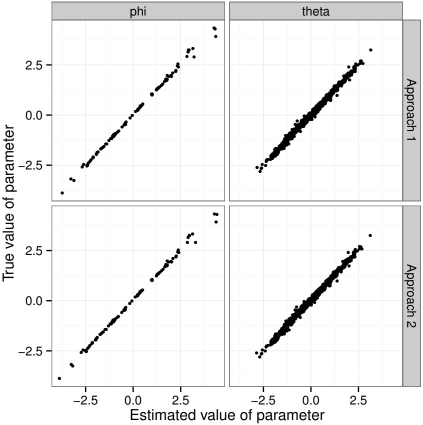
Multiplicative aliasing:

$$= \text{logit}^{-1}\left(\alpha_j + \beta_i - \frac{\gamma}{k^2}((\theta_{it} - \phi_j) \times k)^2\right)$$

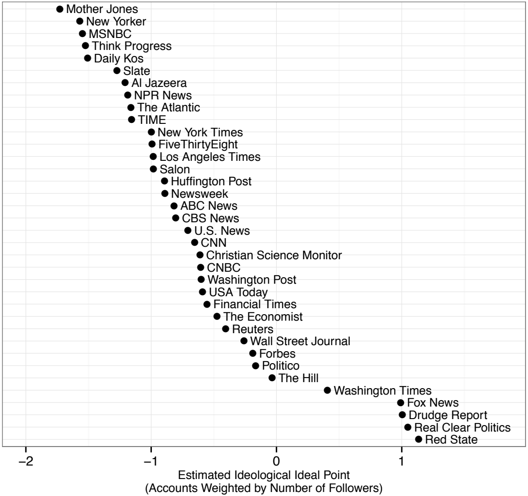
Index

Identifying restrictions

Indeterminacy	Approach 1	Approach 2
Additive aliasing (1)	Fix $\alpha'_j = 0$ or $\beta'_i = 0$	Fix $\mu_\alpha = 0$ or $\mu_\beta = 0$
Additive aliasing (2)	Fix $\phi'_j = +1$ or $\theta'_i = +1$	Fix $\mu_\phi = 0$ or $\mu_\theta = 0$
Multiplicative aliasing	Fix $\phi''_j = -1$ or $\theta''_i = -1$	Fix $\sigma_\phi = 1$ or $\sigma_\theta = 1$

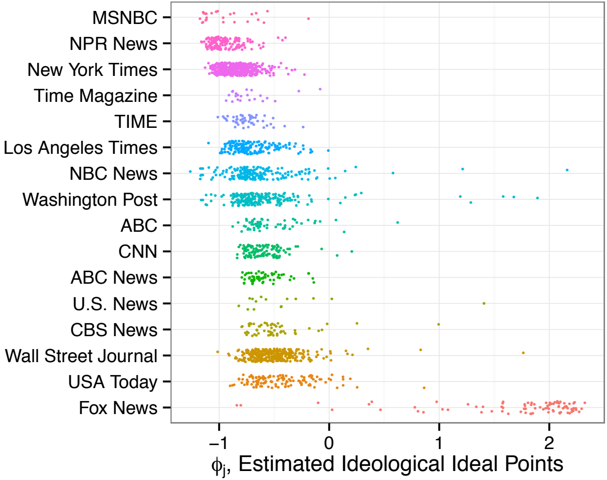


Application: Ideology of Media Outlets and Journalists



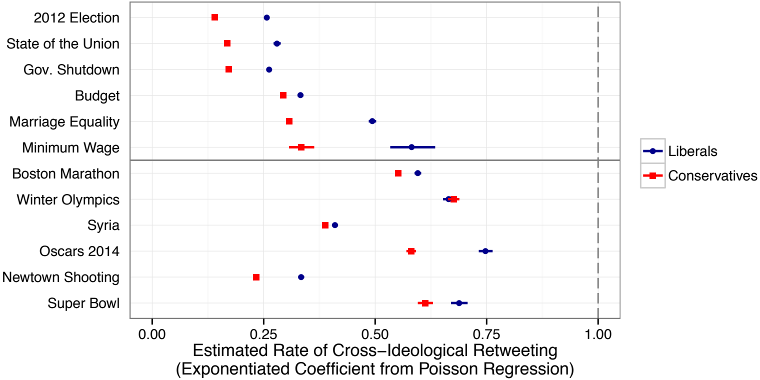
Barberá & Sood (2014) "Follow Your Ideology: A Measure of Ideological Location of Media Sources", *MPSA Conference* Index

Application: Ideology of Media Outlets and Journalists



Barberá & Sood (2014) "Follow Your Ideology: A Measure of Ideological Location of Media Sources", *MPSA Conference* Index

Application: Ideological Asymmetries in Pol. Comm.



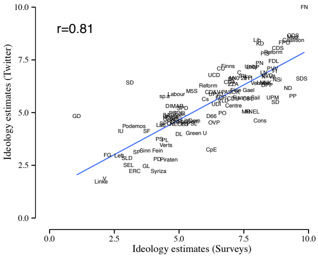
Barberá, Jost, Nagler, Tucker, & Bonneau (2015) "Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?" *Psychological Science* Index

Application: Multidimensional Policy Spaces in Europe

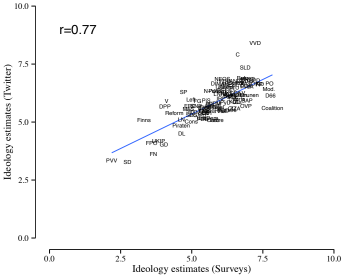
$$P(y_{ij}=1) = \mathrm{logit}^{-1}\left(\alpha_i + \beta_j - \sum_{k=1}^{d} \gamma_d (\theta_{ik} - \phi_{jk})^2\right)$$

Estimated ideological positions for 120 parties in 28 European countries

Left-Right Dimension



Pro/Anti-European Union Dimension



Barberá, Popa, & Schmitt (2015) “Analyzing the Common Multidimensional Political Space for Voters, Parties, and Legislators in Europe”, *MPSA Conference* Index

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

Electoral Institutions and Political Representation

What institutional configurations foster better representation?

Theoretical expectations

Country	Government	Instit.	Congr.	Responsiv.
Germany	Coalition	Prop.	High	Low
Spain	Single-party	Prop.	Medium	Medium
UK	Coalition	Maj.	Medium	Medium
France	Single-party	Maj.	Low	High

Barberá & Bølstad (2015) “A Comparative Study of the Quality of Political Representation Using Social Media Data”, EPSA Conference Paper. Index

Electoral Institutions and Political Representation

j.mp/EPSA-lda-demo

Index