Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data

PABLO BARBERÁ | **Social Media and Political Participation Lab** | **Department of Politics, New York University**

email: pablo.barbera@nyu.edu
web: pablobarbera.com
twitter: @p_barbera

Introduction

- 200M active Twitter users, 500M tweets per day. Can we use this massive amount of data to estimate quantities of substantive interest for social scientists?
- My contribution: a new method that generates (ideology) ideal points estimates for any Twitter user based on which politicians, journalists, and organizations they decide to "follow."
- Relevance: better understanding of representativeness of Twitter population; precise ideal points are required to test hypotheses about political behavior and party competition.
- Key assumption: Twitter users prefer to follow accounts whose ideology is similar to theirs.

→ Why? a) Homophily: clustering in social networks along common traits.
b) Selective exposure: preference for opinion-reinforcing information.

Method

- Users' and politicians' ideology (θ_i and ϕ_j) are defined as latent parameters.
- Data: "following" decisions as a series of binary choices ($\mathbf{Y_{ij}}$).
- Spatial following model: for n users, indexed by i, and m politicians, 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

- Assuming independence, likelihood function is:

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

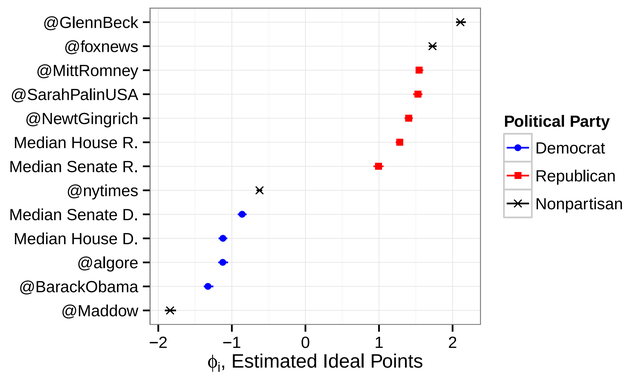
$$\text{where } \pi_{ij} = \alpha_j + \beta_i - \gamma||\theta_i - \phi_j||^2$$

- Model is identified with unit variance restriction on θ.
- The posterior density of each set of parameters is sampled using a Hamiltonian Monte Carlo algorithm (Hoffman and Gelman, 2011).

Data

- $m =$ list of $\sim$1200 highly visible political accounts from US, UK, Spain, Netherlands.
- $n =$ followers of $\geq$3 accounts in each country, collected using Twitter REST API.
- Sample size: 300K in the US, 125K in the UK, 125K in Spain, 96K in the Netherlands.

Estimated Ideal Points for Key Political Actors in the US

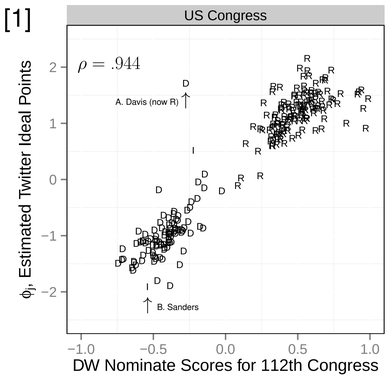


Validation

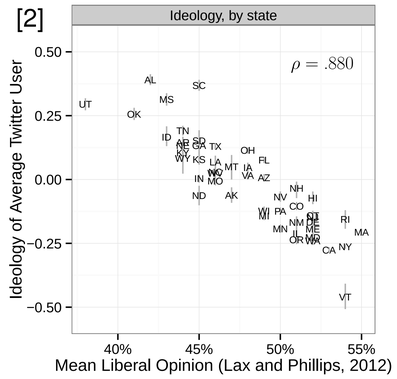
This method is able to correctly classify and scale Twitter users along the left-right dimension. Ideology estimates are highly correlated with...

1. ideal points for members of Congress based on their roll-call votes (DW-NOMINATE)
2. estimates of public opinion by state (Lax and Phillips, 2012)
3. voter registration history in Ohio, after matching with Twitter accounts by full name and county
4. left-right positions in Europe according to expert surveys (Chapel Hill – Bakker et al., 2012)

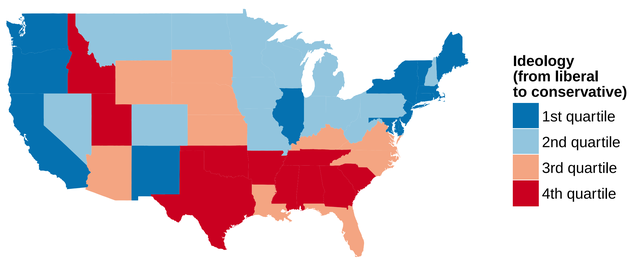
[1]



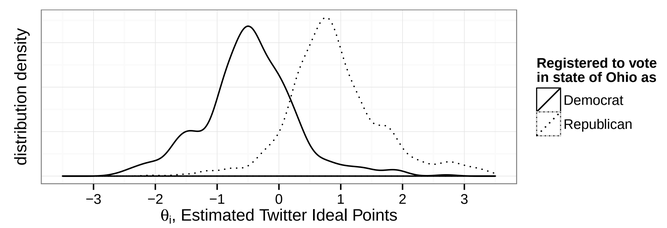
[2]



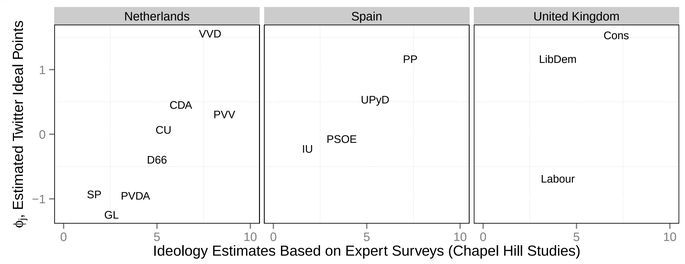
[2]



[3]



[4]

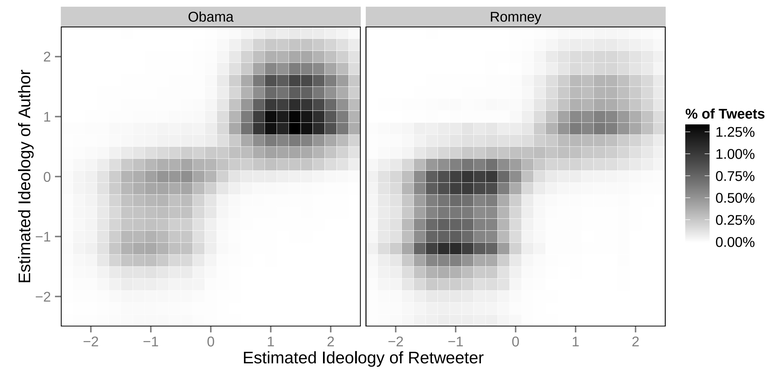


Application

- Social media and political polarization:

→ To what extent is information diffusion through social media bounded by ideology?
→ Is Twitter an "echo chamber" where individuals only receive information that reinforces their existing beliefs or, on the contrary, are cross-ideological interactions frequent?
→ Previous studies find high levels of clustering along party lines (e.g. Conover et al., 2012)

- Data: 75 million tweets mentioning `obama` or `romney` from Aug. 20, 2012 to Nov. 6, 2012 collected using `streamR` package for `R` (Barberá, 2013), available on CRAN.
- Analysis focuses on retweets (indication of information diffusion) sent by a sample of 300K users for which ideology was estimated.
- Results (see figure below):
  1. Structure of conversations on Twitter supports idea of "online echo chamber": most pairs of users involved in retweets mentioning Obama or Romney are ideologically similar.
  2. Differences across ideological groups: liberals tend to engage more often in cross-ideological interactions, while right-leaning Twitter users show more polarized behavior.



Conclusions

- Structure of Twitter networks can be informative about policy positions: an "expert survey" with millions of respondents reporting how ideologically close they perceive politicians.
- Necessary to improve our understanding of how representative Twitter users are.
- Twitter presents several important advantages: vast amount of data, low cost, easy access, real-time nature. Unexplored source of information for old and new research questions.
- Meaningful comparisons of all types of political actors on same ideological scale: useful for wide range of empirical tests that require ideology estimates in Political Science.

References

BAKKER, R., C. DE VRIES, E. EDWARDS, L. HOOGHE, S. JOLLY, G. MARKS, J. POLK, J. ROVNY, M. STEENBERGEN, AND M. A. VACHUDOVA (2012): "Measuring party positions in Europe: The Chapel Hill expert survey trend file, 1999–2010," *Party Politics*.

BARBERÁ, P. (2013): "streamR: Access to Twitter Streaming API via R," R package available on CRAN.

CONOVER, M. D., B. GONÇALVES, A. FLAMMINI, AND F. MENCZER (2012): "Partisan Asymmetries in Online Political Activity," *EPJ Data Science*, 1, 1–19.

HOFFMAN, M. D. AND A. GELMAN (2011): "The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo," *arXiv preprint arXiv:1111.4246*.

LAX, J. AND J. PHILLIPS (2012): "The democratic deficit in the states," *American Journal of Political Science*, 56, 148–166.