OBDA Project

Diogo Mesquita, Etienne Dejoie

NYU, december 2017

---

1. Introduction

The main contribution of [1] consists of a relatively simple method to perform ideal point estimation of legislators and voters in the same ideological scale. And it does so solely based on Twitter's network graph. This is important because other work, such as studies of government formation and stability (for example [2]), require as a first step to know the political ideology of the elements involved in the study. That was previously a difficult scientific problem, which is now more simple to overcome using this method.

Our work presents three separate contributions. First, we replicate the ideology estimate results obtained in [1] for the United States. The data analysis technique we will be using is parameter estimation using sampling methods. In this case, we will use the Hamiltonian Monte Carlo method [3], and the Metropolis-Hasting algorithm [4].

Second, we compare the previous results with the output of a more classic unsupervised method, PCA in this case. Because the task performed in [1] is unsupervised there is no straightforward metric to see how our model performs. Using PCA is useful, therefore, since it provides and additional way to evaluate those results.

Finally, we again apply the method from [1] but this time to France, which was not part of that work's study. Since France's political scene, which includes multiple parties and ideologies, is very different than that of the US, we think this experience provides important information about how this method performs in different political landscapes.

2. Definitions

Throughout this work we define *elites* to be any entity that is relevant from a political standpoint, for example politicians, political parties or some

media outlets, and that is active on Twitter. Additionally, we define *users* to be ordinary citizens that are also politically active on Twitter by means of following elite accounts.

3. Data

3.1. Twitter data

For reproducing the results in [1] for the US, and for applying PCA we simply used the data that was made publicly available from that work. Because the data we gathered has $m = 318$ and $n = 301,537$ elites and users respectively, instead of the reported $m = 548$ and $n = 473,640$ in [1], it is possible that the datasets don't match exactly.

For the France experiment, though, we had to start from the beginning. First, we decided which elite accounts to use. Our choices fell on current politicians, political parties and a set of news outlets. This resulted in $m = 540$ accounts. From these we only selected the ones that were active on twitter. And, because the method to estimate the ideology is very computational expensive, we had to further reduce this number. So we selected only the accounts that had at least 5000 followers in Twitter, resulting in $m = 101$ entities.

These entities combined have $n = 37,514,268$ followers. However, we are not interested in the majority of these, because they either are not from France or are inactive accounts. Furthermore, to improve the quality of the results and restrict the complexity of the model, we did not consider users that had less than 20 tweets, that followed less than 3 elites or that had less than 20 followers. We faced an additional problem, however. Recently, Twitter significantly reduced its API request limits which meant that fetching $n = 37,514,268$ users would take more than one month to do. For that reason, we randomly sampled one out five users from every elite with more than 1,000,000 followers. Despite these cuts, getting the data still took a full week, and we ended up with $n = 43865$ users for France.

4. Model

4.1. The statistical problem

Barbera, in [1] suggests to model the probability that a user follows another user based on three terms:

- the activity of account A (how likely is A to follow an elite)
- the popularity of an elite B (how likely is B of being followed)
- the "political" distance between accounts A and B

Let j be an elite and i be an user. Let y be the adjacency matrix, where y_{ij} equals 1 if user i follows elite j and 0 otherwise. With α_j the popularity of j, β_i the activity of i, θ_i the latent variable of i and ϕ_j, the latent variable of j. The probability of user i following the elite j is:

$$\pi_{ij} = P(y_{ij} = 1|\alpha_j, \beta_i, \theta_i, \phi_j, \gamma) = sigmoid(\alpha_j + \beta_i - \gamma||\theta_i - \phi_j||^2) \quad (1)$$

with γ being some scaling parameter to balance between the two terms of the function.

Assuming that links are independent from each other, i.e. that the decision of any user to follow elites is independent of any other user, the resulting optimization problem we are trying to solve is to maximize the likelihood function:

$$\begin{aligned} p(y|\theta, \phi, \alpha, \beta, \gamma) &= \prod_{ij} p(y_{ij}|\theta_i, \phi_i, \alpha_j, \beta_j, \gamma) \\ &= \prod_i \prod_j sigmoid(\pi_{ij})^{y_{ij}}(1 - sigmoid(\pi_{ij})^{1-y_{ij}} \end{aligned} \quad (2)$$

4.2. The optimization problem

To optimize this problem, we use Markov-Chain Monte Carlo methods, more specifically HCMC. We use gaussian priors, therefore the full posterior join probability is:

$$\begin{aligned} p(\theta, \phi, \alpha, \beta, \gamma|y) = &\prod_i \prod_j sigmoid(\pi_{ij})^{y_{ij}}(1 - sigmoid(\pi_{ij})^{1-y_{ij}} \\ &\prod_j N(\alpha_j|\mu_\alpha, \sigma_\alpha) \prod_j N(\phi_j|\mu_\phi, \sigma_\phi) \\ &\prod_i N(\beta_i|\mu_\beta, \sigma_\beta) \prod_i N(\theta_i|\mu_\theta, \sigma_\theta) \end{aligned} \quad (3)$$

The optimization is done in two steps.

- stage 1: estimate the latent variable for each elite j. We used the pystan library [5], with the No-U Turn algorithm [3]. At this step we estimate α_j and ϕ_j relatives to elites. We also sample parameters relative to users but without updating the underlying variables.

- stage 2: generate samples for the users i. We use the data sampled in stage 1 and complete that with a HCMC [4] hand written in python to sample, in particular, the latent variable θ_i of every user i

For this, we reimplemented the code available from [1] in Python (instead of R). You can find our code on github here.

5. Reproduction of the results with USA dataset

5.1. Our results

First of all, it's worth mentioning that stage 1 took considerably longer to run for us than what is described in [1]. For them it took around 24 hours to run two chains with 1000 iterations, giving them over 500 simulation draws. We couldn't, however, finish this job in less than 7 days in NYU's Prince HPC cluster. So we had to go for less iterations. We fixed at two chains with 150 iterations each, obtaining an effective number of draws of over 50. Which still took more than 3 days to run. We think this is due to the fact that the python version of the Stan library [5] that we used is less efficient than the one implemented in R, which was used in [1], and because in R this library is capable of running multiple chains in parallel while it is not in Python.

Although we have good results, we couldn't match the ones reported in [1]. This is probably due to, as mentioned previously, having a lower number of simulation draws for stage 1. It's even possible that our chains didn't converge as we did not check this.

5.1.1. Elites

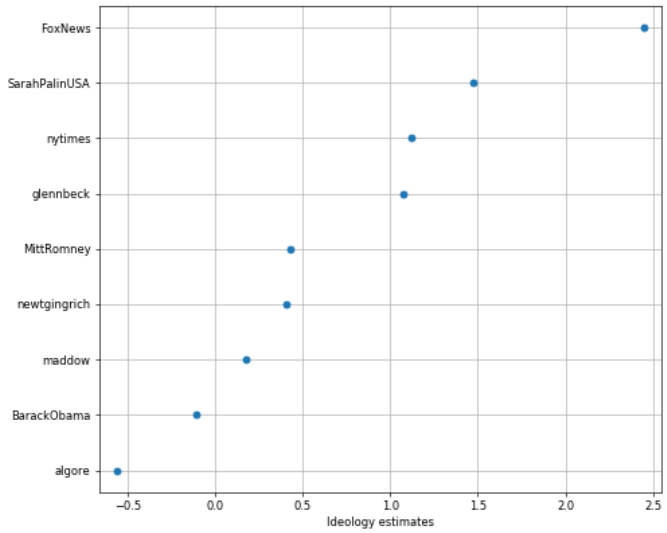

Figure 1: Ideology estimates for key actors in US political scene

Figure 1 shows our model estimations for the ideologies of important political entities of the US. Fox News who has been referred to as a "right-wing propaganda machine" appears, sure enough, with the highest score on the right. In the same side of the spectrum Sarah Palin, who was the Republican party nominee for Vice President in the 2008 election, and conservative Glenn Beck are also assigned ideologies in the range one would expect. On the other side of the scale, previous President Barack Obama and Vice-President Al Gore are both from the Democratic Party and our model correctly places them both on the left. The same for Rachel Maddow, as she is a non-partisan liberal. For the newspaper New York Times, however, our estimation isn't what one would expect. The New York Times is considered to have a liberal bias, among other things because it hasn't endorsed a republican nominee for president since 1956. Therefore it was not expected an ideology estimate of more than 1 like the one we obtained.

In Figure 2 we plot the ideology estimates by grouping the elites by party. From the figure, it is possible to see that the groups have different distributions.

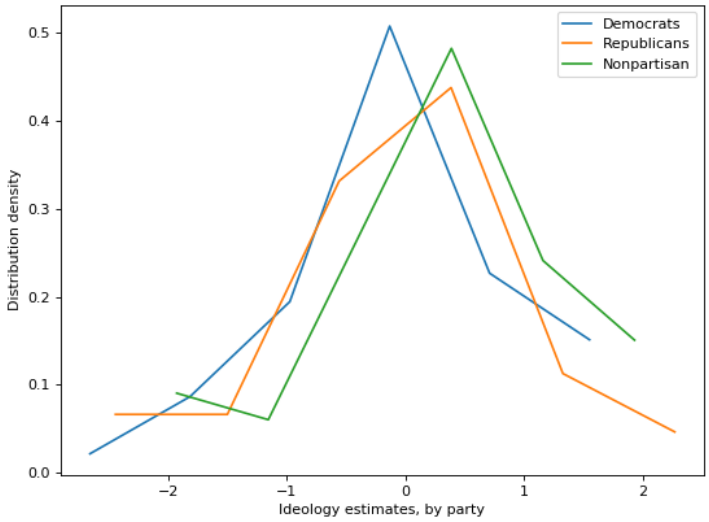

Figure 2: Ideology estimates distribution by party

Concretely, the democrats have on average an ideal point of -0.066 while the republicans are, as expected, to their right with an average of 0.054. The non-partisan group shows up more to the right than both of the previous groups, with an average ideal point of 0.14. We did not expect the results to be so close to 0 for the Democrats and Republicans. Neither did we expect the non-partisan group to be more to the left of the democrats or to the right of the republicans, as was the case. We think this can be explained, again, by the fact that we didn't run the markov chain long enough for it to converge.

5.1.2. Users

Figure 3 shows the distribution of our estimations for the ideal points of elites and ordinary users. On average, political actors are more polarized than the citizens. Which is evidenced by this figure, since a big part of the density of the entities' distribution is on either side of the political spectrum, while the users distribution has most of its density at the middle. Additionally, the (small) majority of users is estimated as liberal, which is also in accordance with the fact that most of the American electorate is liberal (see [6]).

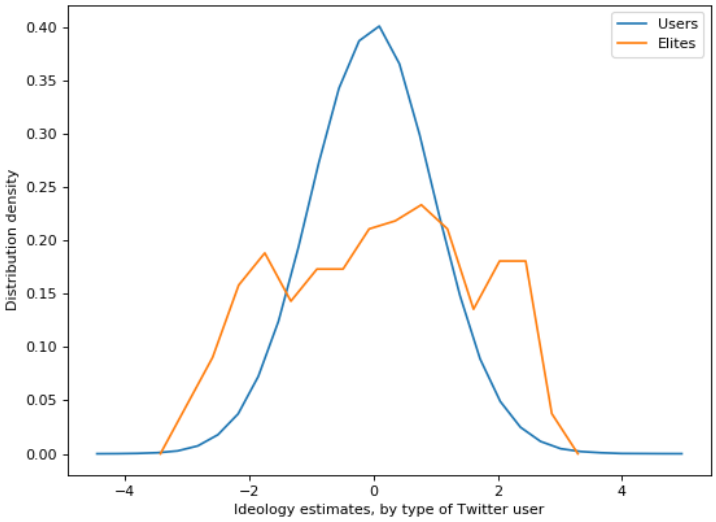

Figure 3: Ideology estimates distribution by account type - Elite or User

This plot, once again, evidences that our ideal point estimation is not as good as in [1]. Concretely because both distributions were expected to be bimodal, which is not the case for either of them.

5.2. Comparison with [1]

Our model performed very close to [1] regarding the ideal points of key political actors. That can be seen in two ways. First, by Figure 1, which is very close to Figure 5 in [1]. Second, by the high correlation of 0.61 between the two estimations.

For the general elites, however, our results weren't so good. In Figure 7 of [1] a clear distinction can be seen between the distributions of elites belonging to different parties. In our case this is only partly true, as can be seen from Figure 2. Moreover, the non-partisan elites distribution is placed between the democrats and republicans in Figure 7 of [1]. In our case, though, it is slightly to the right of the republicans.

As for the results regarding the users, the distribution of ideal points obtained in [1] is bimodal, which is in accordance to previous work (for instance [6]). While our estimations have a normal distribution.

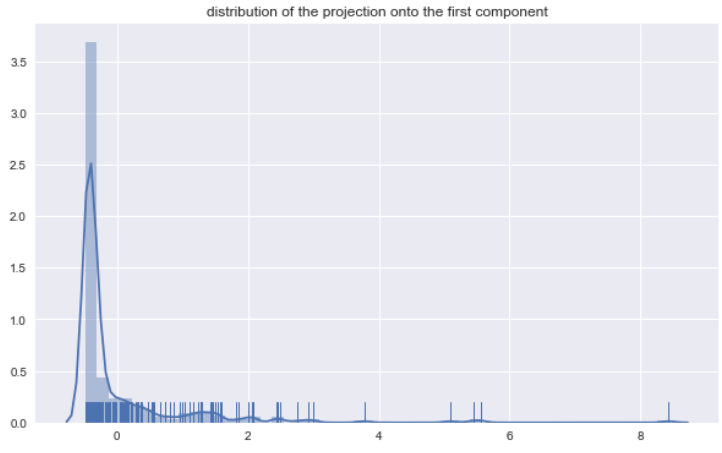

Figure 4: Distribution of coefficient of first component

6. Comparing this method with a PCA

We can leverage the adjacency matrix of connections to perform PCA. In particular, the projection onto the first component gives us a latent variable about each elite. We used this value to compare with the results using the Barbera method.

We see from Figure 4, that the projection onto the first component does not separate the data in different groups. On the opposite, it concentrates all the political actors to a very small area between -0.5 and 2. This justifies the need of a more complex model as the one we explored for the interactions between the political actors.

With no suprise, the correlation and cosine similarity between Barbera's latent variable and the coefficient of the first component are very small, respectively 5% and 4%.

7. Using the model with another dataset: the case of France

We applied Barbera's method to uncover the latent variable of politically active twitter accounts from France. Our goal was to determine if this latent variable would also be a good proxy for the political ideology of these accounts.

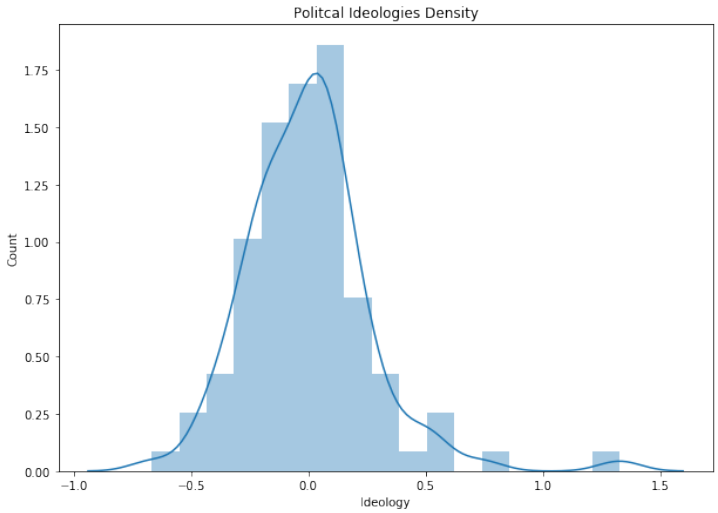

Figure 5: Distribution of the latent variable for the elites

We chose France because it was not studied in [1], because it is a country where Twitter is popular, and because it has a peculiar political scene. The dataset was built as explained in the Data section.

7.1. Distribution of political ideologies of the Elites

Contrary to elites in the US, the distribution of the latent variable for the elites is not bimodal but is a unimodal gaussian, as we can see from Figure 5.

We can understand these results because France political scene is not a bipartite one as in the US. Especially, with the recent election of Emmanuel Macron, France has seen a dramatic change in the former left/right division of French politicians. Now, a lot of politicians who used to belong to different parties are united in a new party which could be qualified as "neither left-wing, neither right-wing".

In addition, Figure 6 shows that our ideal point estimates also cluster the elites by party and that the ideal points actually have meaning. Because REM is a center party, PS a center-left party and LR is center-right, which matches the center of the clusters. FN is an exception, since it is a party on the extreme right, but our estimates place it on the left. However, in the following section we argue why this might actually make sense.

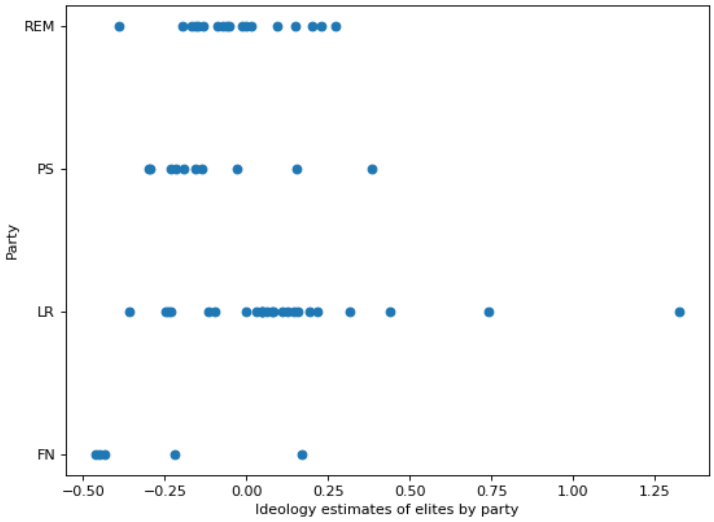

Figure 6: Ideology estimates distribution by party

7.2. Results for key political actors

The first thing we can observe from Figure 7 is that actors that are close politically are also close on this latent variable dimension. For instance, Marine Le Pen (*mlp_officiel*) and Gilbert Collard (*gilbertcollard*) both belong to the Front National.

Interestingly, we can also observe that elites are placed in this ideology dimension in more than a left to right classification. Concretely, anti-system entities are assigned a lower latent variable when compared to entities that "favor" the statu quo. This way, we can see that far-right and far-left parties are united with small latent variables, as Marine Le Pen, Nicolas Dupont Aignan, Gilbert Collard, Jean-Luc Melanchon. On the other side of the spectrum, we get Stephane LeFoll (from Partie Socialiste) and Les Rpublicains, representing the two left and right established parties in French politics.

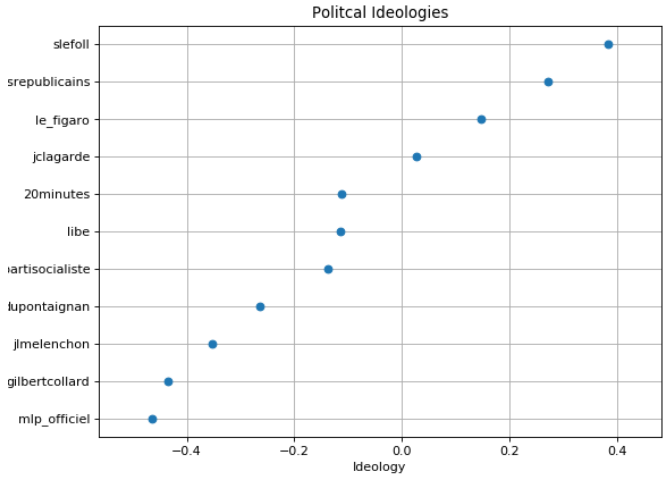

Figure 7: Ideology estimates for key political actors in France

8. Conclusion

Even though we had performance issues regarding the stage 1 of the method, our work seems to corroborate the results obtained in [1]. The ideal point estimates of key political actors from the United States, for instance, align almost perfectly, in relative terms, with their actual political ideologies.

In the case of France, this method is capable of not only classifying differently political actors depending on their political organization, but also gives us an interesting insight into French politics.

We also concluded that naively applying a simple unsupervised method like PCA to this problem does not produce ideal point estimates.

9. References

[1] P. Barberá, Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data, Political Analysis 23 (2014) 76–91.

[2] M. Laver, K. A. Shepsle, Making and breaking governments: Cabinets and legislatures in parliamentary democracies, Cambridge University Press, 1 edition, 1996.

[3] A. Gelman, A. J. Carlin, H. Stern, D. Rubin, Bayesian data analysis, Chapman and Hall/CRC, 3 edition, 2013.

[4] M. C. Herron, J. Bafumi, Equation of state calculations by fast computing machines, J Chem Phys 21 (1953) 519–42.

[5] Stan, Development, Team, Pystan: the python interface to stan, `http://mc-stan.org`, 2017.

[6] M. C. Herron, J. Bafumi, Leapfrog representation and extremism: A study of american voters and their members in congress, American Political Science Review 104 (2010) 519–42.