Estimating Political Ideology using (massive) Social Network Data

Project Report - MACS 30123

Sudhamshu Hosamane, Koichi Onogi

May 25, 2023

Introduction

Political ideology plays a crucial role in shaping policy decisions and electoral outcomes. Estimating policy positions is critical for analyzing political behavior and policy outcomes. Political actors are often motivated by their ideological beliefs when making decisions about policies and issues. By estimating the ideological positions of politicians and voters, we can gain insights into their likely behavior and decision-making. This, in turn, can inform policy analysis and improve the effectiveness of policy-making

Learning the ideological leanings of a politician is important for voters, as it helps them make informed decisions at the ballot box. Voters are more likely to support a candidate whose policy positions align with their own beliefs. By understanding the ideological leanings of a politician, voters can make informed decisions about which candidate to support and which policies to advocate for.

Similarly, learning the ideological leanings of voters is important for politicians, as it helps them understand the preferences and priorities of their constituents. By understanding the ideological leanings of voters, politicians can tailor their policies and campaign messages to appeal to their base and build support for their platform.

Although this complex and multifaceted concept has long been of interest to scholars and practitioners alike, there hasn't been a reliable way to operationalize policy positions. Researchers have been using qualitiative methods like analyzing congressional voting records, analyzing speeches made by these legislators and conducting interviews with them or their stakeholders who are familiar with their policy positions. These methods have been resource intensive, erroneous and are not scalable to a wider audience. The advancement of methodological approaches in Political Science have developed methods to estimate ideology as a continuous latent variable for measuring policy position (Poole and Rosenthal 1999), (Bonica 2013). While these methods overcome issues associated with qualitative analysis, the data used in these methods are often incomplete (campaign contributions) (Bonica 2013), sparse (Poole and Rosenthal 1999) and generate estimates that are static in the short run. Most of these methods are studied only for legislators or self-selected voters' sample and is infeasible to gather complete data (using surveys) for a large sub-population.

Under the assumption that social networks are homophilic, researchers have users' position in a social network as a source of information about their ideological positions (Barbera, Preeti). Following a similar approach, in this project, I attempt to estimate Political Ideology using Twitter

follower network data for 535 members of the 116th Congress, and their Twitter followers (n = 3,000,000) using a Bayesian Ideal Point Estimation model.

Estimating Ideology as a Latent Variable

Since the pioneering work of Poole and Rosenthal (1999), a number of scholars have used spatial voting models to estimate ideological preferences from rollcall votes and other data in the fields of comparative politics and international relations as well as American politics [refs]. With the increasing availability of data and methodological sophistication, researchers have recently turned their attention to the estimation of ideological leanings that are spatially and temporally comparable. Bailey (2007) measures ideal points of U.S. presidents, senators, representatives, and Supreme Court justices on the same scale over time. Researchers have also begun to analyze large data sets of more complex information. For example, Slapin and Proksch (2008) develop a statistical model that can be applied to estimate ideological positions from textual (syntactic) data using occurence of words as a cue.

Researchers in Computer Science employed advancements in Topic modelling and Neural Networks to include semantic information in tweets to predict ideology estimates. Gerrish and Blei (2011) predict the voting patterns of Congress members based on supervised topic models. Iyyer et al. (2014) use recurrent neural networks to predict political ideology of congressional debates and articles in the Ideological Book Corpus (IBC) and demonstrate the importance of compositionality (modifier phrases, punctuation) in predicting ideology. Kulkarni et al. (2018) incorporate attention mechanism in their neural network model to capture cues from both textual content and network structure of news articles to estimate ideology positions of users.

More recently, Lahoti, Garimella, and Gionis (2018) model the problem of learning the liberal-conservative ideology space of social media users and media sources as a constrained non-negative matrix-factorization problem using Twitter follower network data and content-consumption information. Similarly, previous studies have used Bayesian latent space models applied to social networks to allow for the estimation of ideal points for a massive scale of individuals along with error in estimation (Hoff, Raftery, and Handcock 2002), (Barberá 2015).

Assumptions

Estimating political ideology using Twitter follower-network data requires some strong bridging assumptions about user behaviour -

- Twitter Politician-follower networks are homophilous, i.e, Twitter users follow legislators whose ideology aligns with their own.

  These assumptions are similar to the ones made in other studies that use Twitter follower network data to estimate ideology of users.

- The cost to follow a legislator on Twitter is not trivial. Following a political elite might be associated with a considerable cost -

  Tweets by elites might cause a cognitive dissonance for users whose ideologies are not aligned with users. Such tweets might prompt the user to unfollow the elite.

  Twitter users have limited amount of time and attention to spend on Twitter. This creates an opportunity cost since it reduces the likelihood of being exposed to other messages.

Data

I identified the legislators of the 116th U.S Congress as the Twitter elites for my study. The 116th Congress had

- 100 senators, two from each of the fifty states.
- 435 representatives, seats are distributed by population across the fifty states.
- 6 non-voting members from the District of Columbia and US territories which include American Samoa, Guam, Northern Mariana Islands, Puerto Rico, and US Virgin Islands.

Rauhauser (2019b), Rauhauser (2019a) assimilated a directory of files in 2019 that contains followers (as Twitter IDs) corresponding to each Twitter handle (which is the name of each file). Apart from Official (Office) Twitter handles, these folders also contained handles (along with the list of followers) for several legislators' campaign and personal Twitter accounts. To ensure that I only retrieved information for handles corresponding to official accounts, I used the official CSPAN twitter handle dataset curated by Siddique (2019), and retrieved the list of followers for this subset and stored the network as an adjacency list (of directed edges from legislators to all their unique followers). This resulted in a total of 535 legislators with 6 missing in all - Rep. Collin Peterson (@collinpeterson), Rep. Greg Gianforte (@GregForMontana), and Delegate Gregorio Sablan (@Kilili_Sablan), Sen. Rick Scott (@SenRickScott), and Delegate Michael San Nicolas (No Twitter account).

Since some Twitter elites also followed each other, to build a bipartite network I retrieved the Twitter IDs of all 535 legislators using the dataset collected by Wrubel and Kerchner (2020) and removed these IDs from the follower lists. To compare my final Twitter ideal point estimates to expert ratings, I mapped each of the twitter elite to their DW-NOMINATE scores, published by Voteview (J. B. Lewis et al. 2023). The total number of unique twitter users (followers) with this subset of elites were 16,420,157. These followers ranged from following only 1 Twitter elite to all 535 Twitter elites. Users who followed too few or too many legislators would not provide much information about the ideology of themselves or the elites they are following. To mitigate this issue, I removed all users who followed less than 3 or more than 300 legislators. This subset the final set of Twitter followers to 3,962,197 people. I sampled 3,000,000 users from this subset (uniformly random) to create a final adjacency matrix (of directed edges from oridnary users to elites) of size [3,000,000 $\times$ 535] (Users $\times$ Legislators) .

Model

To estimate the ideology of both Twitter elites (legislators) and ordinary users, I used a latent space model applied to social networks similar to the one described by Hoff, Raftery, and Handcock (2002). More specifically, I follow the model described by Barberá (2015), with a few adjustments to the parameters. There are N (N=3,000,000) ordinary Twitter users who may or may not choose to follow any of the M Twitter elites (M=535). Suppose i is an ordinary Twitter user such that i $\in$ {1,2,...,N} and j is a Twitter elite such that j $\in$ {1,2,...,M} and let $y_{ij} = 1$ if i decides to follow j and 0 otherwise. From our previous assumptions, this directed link exists if the ideological positions of both the user and the elite are close to each other. Let us consider the ideology estimates of a user and an elite to be θ_i and w_j respectively, such that θ_i, $w_j \in \mathbb{R}$. Let us consider two additional latent variables α_i and $\beta_j \mid \alpha_i$, $\beta_j \in \mathbb{R}$, to account for observed confounding - a Twitter elite's popularity and an ordinary user's interest in Politics respectively. Higher values of α and β suggest more popularity and higher interest in Politics respectively. Therefore, we can model the probability

of a user following an elite as -

$$P(y_{ij} = 1 \mid \alpha, \beta, w, \theta, \gamma) = \sigma(\alpha_j + \beta_i - \gamma||\theta_i - w_j||^2)$$

where γ is a normalising constant and σ is the inverse of the logistic function. Since none of $\theta_i, \beta_i \mid i \in \{1, .., N\}$ and $\alpha_j, w_j \mid i \in \{1, .., M\}$ are observed directly, the task is to estimate these latent variables for all users and elites. Assuming local independence (individual decisions to follow are independent across users N and M, conditional on the estimated parameters), this can be formulated as a problem to maximise the likelihood of Bernoulli variable y^* where $y^*_{ij} = \alpha_j + \beta_i - \gamma||\theta_i - w_j||^2$ -

$$p(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \sigma(y^*_{ij})^{y_{ij}} (1 - \sigma(y^*_{ij}))^{1-y_{ij}} \tag{1}$$

Barberá (2015) explains that the above model is unidentifiable - any constant can be added to all the parameters θ_i and w_j without changing the predictions of the model; and similarly θ_i and w_j can be multiplied by any nonzero constant, with γ divided by its square, leaving the model predictions unchanged. In order to constrain these estimates, I modeled an informative prior with a unit variance for θ, and non informative Normal priors for the rest of the parameters. With this, the full joint posterior distribution is -

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\theta}, \gamma \mid \boldsymbol{y}) \propto \prod_{i=1}^{N} \prod_{j=1}^{M} \sigma(y^*_{ij})^{y_{ij}} (1 - \sigma(y^*_{ij}))^{(1-y_{ij})} \prod_{j=1}^{M} \mathcal{N}(\alpha \mid \mu_\alpha = 0, \sigma_\alpha) \tag{2}$$
$$\prod_{j=1}^{N} \mathcal{N}(\beta \mid \mu_\beta, \sigma_\beta) \prod_{j=1}^{M} \mathcal{N}(w \mid \mu_w, \sigma_w) \prod_{j=1}^{N} \mathcal{N}(\theta \mid \mu_\theta = 0, \sigma_\theta = 1)$$

Estimation of ideal points using MCMC samples

To hasten the convergence process, I followed the implementation recommendation by Barberá (2015). I set the starting values for α as the logarithm of the number of followers corresponding to the elite, β as the logarithm of the number of elites the corresponding user follows, w as -0.5 for legislators from the Democratic Party, +0.5 for legislators from the Republican Party and 0 for legislators with no party affiliation (independent). I used the No U-Turn Sampler available (Homan and Gelman 2014) in the RStan Package to get the values for $\sigma_\alpha, \mu_\beta, \sigma_\beta, \mu_w$ and σ_w and to get the starting values for other estimators. In the second stage I used a Metropolis-Hastings algorithm (Metropolis et al. 1953) to draw samples from the posterior distribution starting from the values obtained using RStan. 1000 iterations of running the MCMC algorithm took longer than 20.5 hours and was met with memory overflow issues.

Estimation of ideal points using variational-EM

Since it wasn't feasible to estimate ideal points using the whole dataset with an MCMC algorithm (or its variants) I attempted to use the Expectation-Maximization algorithm for the posterior described in eq'n (2) using the approach described by Imai, Lo, and Olmsted (2016). The authors derive the joint variational distribution for the EM algorithm for the ideal point model of interest under the following factorization assumption -

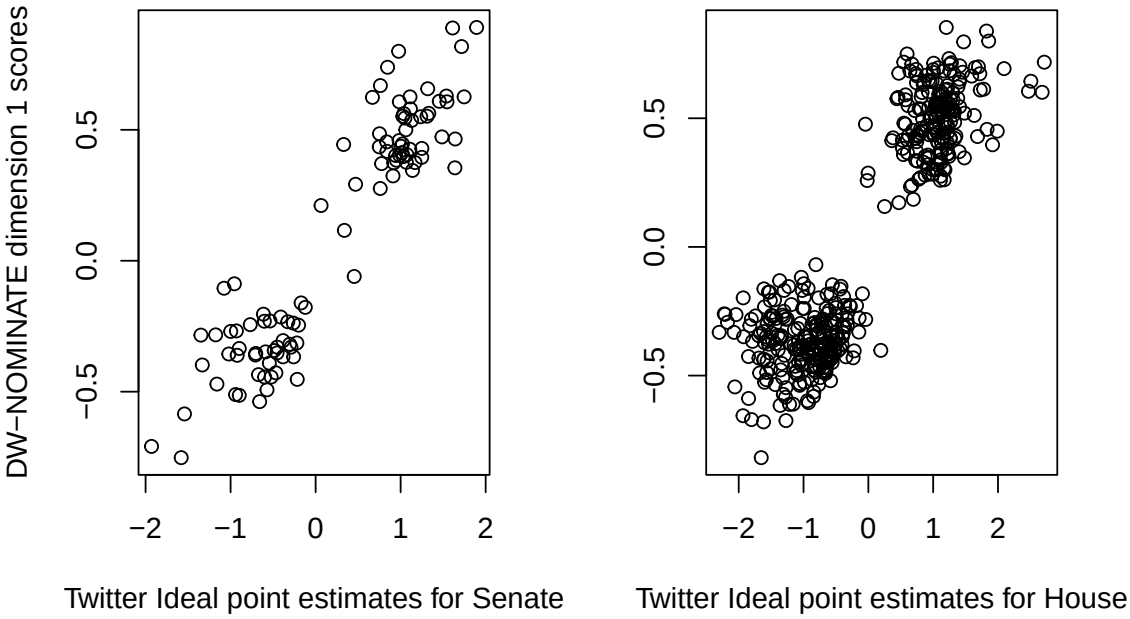
$$q(\mathbf{y}^*, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j=1}^{M} q(y^*_{ij}) \prod_{i=1}^{N} q(\beta_i) q(\theta_i) \prod_{j=1}^{M} q(\alpha_j)(w_j)$$

where the conditional variational distributions $q(y^*_{ij} \mid \beta_{i-1}, \theta_{i-1}, \alpha_{j-1}, w_{j-1})$, $q(\beta_i \mid y^*_{ij}, \theta_{i-1}, \alpha_{j-1}, w_{j-1})$, $q(\theta_i \mid y^*_{ij}, \beta_i, \alpha_{j-1}, w_{j-1})$, $q(\alpha_j \mid y^*_{ij}, \beta_i, \theta_i, w_{j-1})$, and $q(w_j \mid y^*_{ij}, \beta_i, \theta_i, \alpha_j)$ are derived in the supplementary appendix of Imai, Lo, and Olmsted (2016).

Imai, Lo, and Olmsted (2016) implemented their proposed EM algorithm in the `emIRT` package (Imai, Lo, and Olmsted (2022)), available through the Comprehensive R Archive Network. I utilised this package for the estimation of the ideal points for the entire dataset (3,000,000 users and 535 elites). I set up informative priors for all 4 latent variables - $\alpha \sim \mathcal{N}(0(mean), 25(variance)), \beta \sim \mathcal{N}(0, 25), \theta \sim \mathcal{N}(0, 1)$ and $w \sim \mathcal{N}(0, 1)$ and ran the EM algorithm on a HPC node with 28 cores (1 thread per core) for 5000 iterations. This execution took about 50 hours. I used the same starting values for all latent variables as described for [Model 1]. I ran 1000 iterations at a time and used the values obtained from that run as the starting values for the next. I noticed that estimated values for all the variables of interest changed only slightly from 4000th iteration to 5000th iteration (change in the order of 10^{-3}), and the standard errors for $\boldsymbol{\theta}$ and $\boldsymbol{w}$ were of the order 10^{-4} and 10^{-7} on average respectively, suggesting that the values were close to convergence.

Results and Validation

Analysis of Legislators' and Users' Ideal Points



The above figure compares w_js, the ideal point estimates, of 535 members of the 116th US Congress based on their Twitter network of followers (x-axis) with their DW-NOMINATE scores based on their roll call voting records (Poole and Rosenthal 1999), on the y-axis. We can see that the estimated ideal points are clustered into two different groups that align almost perfectly with party membership. The correlation between Twitter and roll-call-based ideal points is 0.903 in the House and 0.932 in the Senate. This is comparable with the Twitter estimates for the 112th Congress ($\rho =$ 0.941 in the House and 0.954 in the Senate) estimated in (Barberá 2015) using MCMC sampling. The correlation between Twitter estimates and DW-NOMINATE scores for Senate-Democrats, Senate-Republicans, House-Democrats and House-Republicans are 0.527, 0.564, 0.121 and 0.327 respectively.

Let us try to analyse some of these estimates qualitatively.

Let us check the 5 most left-leaning legislators according to DW-NOMINATE scores (along with their scores) -

```
data$names[which(data$nominate_dim1 %in% head(sort(data$nominate_dim1),5))]
```

```
## [1] "kamala harris"    "elizabeth warren" "barbara lee"      "sean casten"
## [5] "sylvia garcia"
```

```
data$nominate_dim1[which(data$nominate_dim1 %in% head(sort(data$nominate_dim1),5))]
```

```
## [1] -0.709 -0.752 -0.680 -0.675 -0.819
```

Although these are well known Democrats, none of these are radically left leaning. Senator Bernie Sanders of Vermont has consistently advocated for policies that are further to the left than those advocated by Kamala Harris and Elizabeth Warren. Sanders has called for universal healthcare, free public college, and higher taxes on the wealthy, which are all more progressive policies than those Harris and Warren have proposed.

In terms of voting records, Congresswoman Alexandria Ocasio-Cortez has a higher progressive score than Barbara Lee, Sean Casten, and Sylvia Garcia according to Progressive Punch ("ProgressivePunch: House Members by Score / All Issues. ProgressivePunch" 2023), a website that tracks voting records and legislative actions. Ocasio-Cortez has been a vocal advocate for policies such as the Green New Deal and Medicare for All, which are considered more left-leaning than some of the policies supported by Lee, Casten, and Garcia.

While media coverage of New York Congresswoman Alexandria Ocasio-Cortez describes her as liberal or even *ultra-liberal* for supporting the green new deal, reparations for slavery, and abolishing the Immigration and Customs Enforcement (ICE) – all positions associated with the left wing of the Democratic caucus, surprisingly NOMINATE scores rank her as a Moderate. J. Lewis (2022) argues that while there are votes on which Ocasio-Cortez joins other strong liberals in voting against the moderates in her party, there are a handful of votes on which Ocasio-Cortez has gone against nearly every other Democrat including Omar, Tlaib and Underwood and sided with (nearly) every Republican. While Congressional observers understand that votes like Ocasio-Cortez's vote against the rules of the House are protest votes, NOMINATE does not. Rather, NOMINATE sees these as instances in which Ocasio-Cortez looks like a conservative and it adjusts her location to be more conservative accordingly.

Let us now check the 5 most left-leaning legislators using our estimated Twitter Ideal Points -

```
data$names[which(results$means$w %in% head(sort(results$means$w),5))]
```

```
## [1] "alexandria ocasio-cortez" "ayanna pressley"
## [3] "deb haaland"              "ilhan omar"
## [5] "rashida tlaib"
```

```
results$means$w[which(results$means$w %in% head(sort(results$means$w),5))]
```

```
## [1] -2.223 -2.301 -2.072 -2.173 -2.211
```

We notice that the above legislators (barring Deb Haaland) are part of '*The Squad*', a progressive group of Democratic legislators that has gained significant attention in recent years for their vocal support of progressive policies and criticism of the political establishment. All of these are women of color, come from working-class backgrounds, which has influenced their policy priorities and advocacy for policies that benefit working-class Americans, have all been involved in grassroots activism and advocacy prior to their election to Congress, and they all have been vocal advocates of Medicare for All, criminal justice reform, immigration reform and the 'The Green New Deal'.

Let us identify the 5 most conservative legislators based on DW-NOMINATE scores

```
data$names[which(data$nominate_dim1 %in% tail(sort(data$nominate_dim1),5))]
```

```
## [1] "rand paul"    "ted cruz"     "mike lee"     "andy biggs"   "ralph norman"
```

```
data$nominate_dim1[which(data$nominate_dim1 %in% tail(sort(data$nominate_dim1),5))]
```

```
## [1] 0.889 0.818 0.891 0.839 0.853
```

and using Twitter Ideal Point estimates.

```
data$names[which(results$means$w %in% tail(sort(results$means$w),5))]
```

```
## [1] "jim jordan"     "john ratcliffe" "louie gohmert"  "mark meadows"
## [5] "matt gaetz"
```

```
results$means$w[which(results$means$w %in% tail(sort(results$means$w),5))]
```

```
## [1] 2.714 2.094 2.507 2.471 2.681
```

All these legislators (except Representative John Ratcliffe) are/were members of the Freedom Caucus, also known as the House Freedom Caucus, a congressional caucus consisting of Republican members of the United States House of Representatives. It is generally considered to be the most conservative and farthest-right bloc within the House Republican Conference (Desilver 2015).

The caucus is positioned right-wing to far-right on the political spectrum, with some members holding right-wing populist beliefs such as opposition to immigration reform. The group takes hardline conservative positions and favors social conservativism and small government. The group has also sought to repeal the Affordable Care Act many times. Jim Jordan and Mark Meadows were the first and the second chairs of the caucus respectively. John Ratcliffe an attorney who served as the Director of National Intelligence from 2020 to 2021 was regarded as one of the most conservative members in the 116th Congress (Dilanian 2019).

Now let us check the legislators who are most neutral (or bipartisan).

5 most neutral legislators using DW-NOMINATE scores and the distance of their estimates from 0 -

```
data$names[which(abs(data$nominate_dim1) %in% head(sort(abs(data$nominate_dim1)),5))]
```

```
## [1] "doug jones"     "kyrsten sinema" "susan collins"  "joe manchin"
## [5] "ben mcadams"
```

```
head(sort(abs(data$nominate_dim1)),5)
```

```
## [1] 0.060 0.069 0.088 0.105 0.116
```

Sorting the 5 most neutral/bipartisan legislators, and their distance from 0 using the estimated Twitter Ideal Points -

```
data$names[which(abs(results$means$w) %in% head(sort(abs(results$means$w)),5))]
```

```
## [1] "lisa murkowski"     "ed perlmutter"      "jenniffer gonzalez"
## [4] "jim hagedorn"       "will hurd"
```

```
head(sort(abs(results$means$w)),5)
```

```
## [1] 0.00312 0.01777 0.04020 0.04485 0.06411
```

Learning that Sen. Murkowski is often described as one of the most moderate Republicans in the Senate and a crucial swing vote [GovTrack (2020b)](Raju and Goode 2011) is a strong validation for the credibility of the estimated ideal point values of the bipartisan politicians. Representing the Republican Party in the Senate, she voted with President Barack Obama's position 72.3% of the time in 2013, one of only two Republicans voting for his positions over 70% of the time. In recent years, she opposed Brett Kavanaugh and supported Ketanji Brown Jackson in their respective nominations to the Supreme Court.

As of August 2019, Hurd was the only black Republican in the House of Representatives. During his congressional tenure, Hurd was known for his expertise in technology and cybersecurity as well as for his bipartisanship (GovTrack 2020a). According to a 2016 Vote Smart analysis, he supported anti-abortion legislation, opposed an income tax increase, supported building the Keystone Pipeline, opposed the federal regulation of greenhouse gas emissions, opposed gun-control legislation, and supported repealing the Affordable Care Act.

Now, let us try to find 5 legislators whose ideologies are closest to any given legislator (Rep. Adam Schiff as an example) -

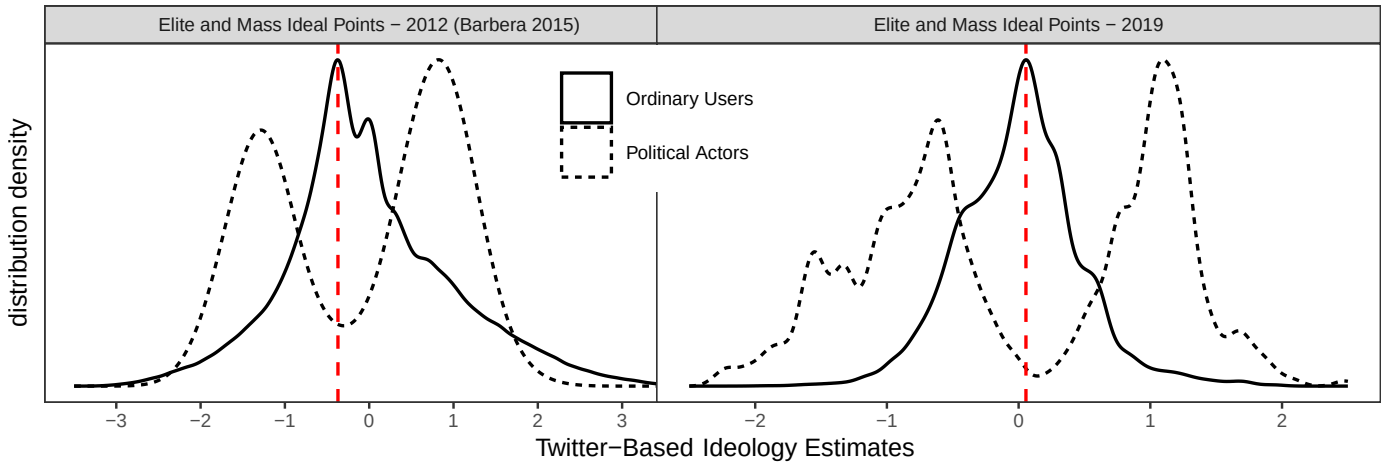
```
find_k_closest_to('adam schiff',5)
```

```
## [1] "nanette barragan" "elizabeth warren" "hakeem jeffries"  "elissa slotkin"
## [5] "colin allred"
```

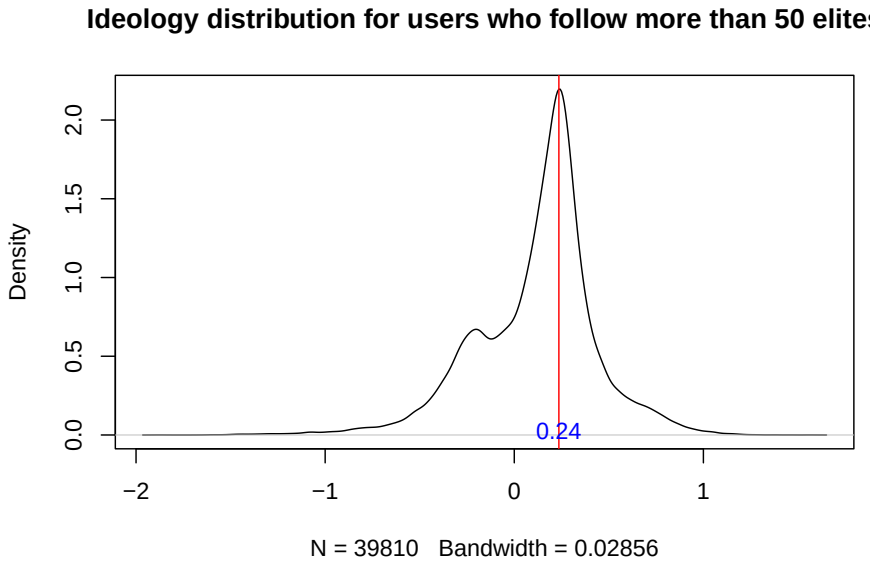
We see that there are many similarities between the policy positions of Rep. Adam Schiff and Rep. Nanette Barragan. For example, both Schiff and Barragan have been vocal advocates for healthcare reform and have supported policies such as the Affordable Care Act and the creation of a public option for health insurance. They have also both been strong supporters of immigrant rights, and have opposed the Trump administration's efforts to restrict immigration and roll back protections for undocumented immigrants. Additionally, they have both supported measures to strengthen gun control laws and reduce gun violence.

Analysis of Mass Ideology at the Aggregate Level

Further, I compare the mass ideology distributions between Twitter users in 2012 vs Twitter users in 2019.



We can observe that the mode of the mass ideology distribution has shifted from being slightly liberal (-0.371) in 2012 to neutral (0.0556) in 2019, suggesting that a lot more people who were right-leaning joined or became more active on Twitter since 2012. Now let us check the distribution for users who are relatively more interested in politics (who follow more than 50 legislators)



We see that users (or bots) who are more interested in politics (by following more than 50 legislators) are more inclined towards the Conservative ideology. Let us try to find the relation between β (latent variable for interest on politics) and the Twitter ideology ideal point estimates (θ)

```
df <- data.frame(theta = as.numeric(results$means$theta), beta = as.numeric(results$means$beta))
lm_robust(theta~beta, data=df)
```

```
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper    DF
## (Intercept) 0.004477  0.0002543   17.61 2.248e-69 0.003978 0.004975 3e+06
## beta        0.207031  0.0005649  366.47 0.000e+00 0.205924 0.208138 3e+06
```

The regression coefficients strengthens our assumption from the previous plot that people more interested in politics are on average right-leaning. The 95% C.I for the regression estimator for β is very narrow - (0.205, 0.208), suggesting that people showing more interest in politics on Twitter tend to be supporting the Republican party. We also infer from the regression results that a person with almost no political interest ($\beta_i = 0$) is non-partisan or neutral ($E[\hat{\theta}_i] \approx 0$).

Let us validate our model further by evaluating the probability of tie between an average Twitter user and a legislator (Sen Kamala Harris as an example).

The means $\bar{\beta}$ and $\bar{\theta}$ respectively are -

```
mean_beta <- mean(as.numeric(results$means$beta))
mean_beta
```

```
## [1] -0.0001695
```

```
mean_theta <- mean(as.numeric(results$means$theta))
mean_theta
```

```
## [1] 0.004442
```

There for our probability estimate turns out to be (assuming $\gamma = 1$) -

$$P(\hat{y}_{Avg\,user-K.\,Harris} = 1) = \sigma(\alpha_{K.\,Harris} + \bar{\beta} - ||\bar{\theta} - w_{K.\,Harris}||^2)$$

```
## [1] "The number of Twitter followers for kamala harris are 404283"
```

```
## [1] "The probability of an average Twitter user following kamala harris is 0.1504"
```

```
## [1] "The number of followers estimated for kamala harris using the assumed mode are 451332"
```

We can see that the actual number of followers in the subset data (404283) is not too far off from the estimated value - 451332.

Let us try estimating the followers for another legislator - Sen. Bernie Sanders:

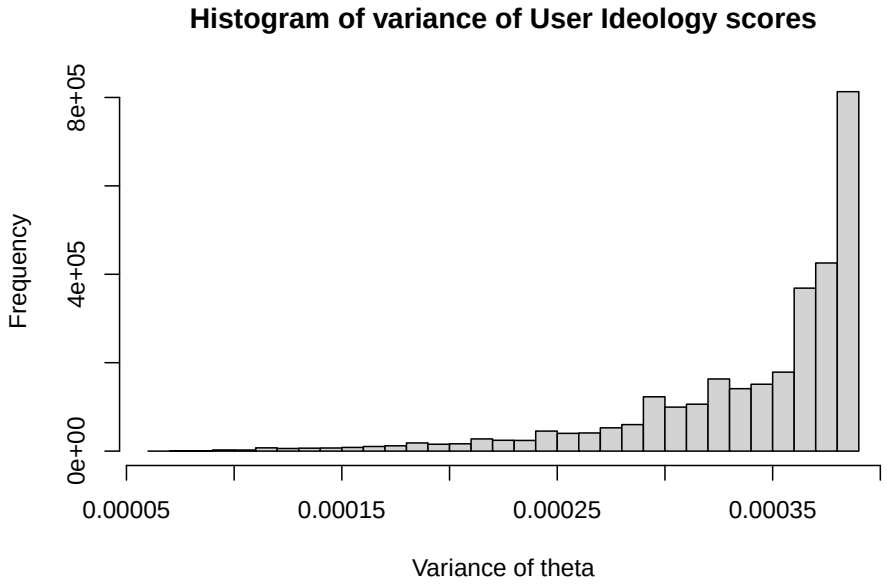
```
## [1] "The number of Twitter followers for bernie sanders are 934699"
```

```
## [1] "The probability of an average Twitter user following bernie sanders is 0.3861"
```

```
## [1] "The number of followers estimated for bernie sanders using the assumed mode are 1158247"
```

The actual number of followers (in the subset data) and the estimated number of followers are comparable (of the same order) for Senator Bernie Sanders too.

Finally, let us analyse our estimates with respect to their standard errors - Plotting the standard errors of latent variable θ -



We see that there is a high number of ideal points with a variance larger than 0.000375. Let us try to elicit more information -

```
summary(rowJ[which(results$vars$theta>0.00037)]) #rowJ represents the number of elites a user J is foll
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    3.00    5.00    9.44   10.00  300.00
```

We see that users whose ideal points have a variance over 0.00037 follow anywhere between 3 and 300 elites with a median of 5 elites. This doesn't reveal much information. Now let us try to analyse the magnitude of the ideal points for the variances of interest

```
summary(abs(results$means$theta[which(results$vars$theta>0.00037)]))
```

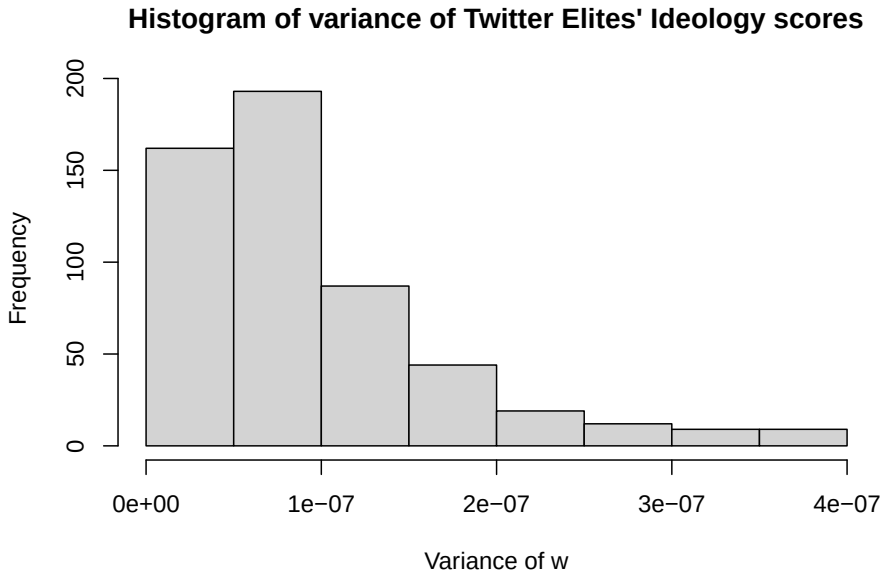
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0463  0.0970  0.1020  0.1534  0.2358
```

```
summary(abs(results$means$theta[which(results$vars$theta<0.00037)]))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.210   0.310   0.455   0.526   0.626   2.431
```

It is evident that ideal points that are closer to 0 have a higher variance (most of the variances for θ<0.2 is greater than 0.00037) as compared to ideal points that are away from 0 ($|\theta|$>0.25).

Let us analyse standard errors for the ideal points of Twitter elites.




```
summary(colK[which(results$vars$w<10**(-7))]) #colK represents number of followers for elite k
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     760    6824   12859   47711   31280  996687
```

```
summary(colK[which(results$vars$w>20**(-7))])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     304    6824   13304   46962   30952  996687
```

```
summary(abs(results$means$w[which(results$vars$w<10**(-7))]))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.792   1.003   1.165   1.244   1.409   2.714
```

```
summary(abs(results$means$w[which(results$vars$w>20**(-7))]))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0031  0.6772  0.9999  1.0057  1.2656  2.7141
```

However, we do not observe any such corresponding patterns for the variance of w (similar to the variances of θ)

Conclusion and future work

With all these observations from the previous section, we find a strong evidence that online social networks can serve as a rich source of information about a person's ideological position. I show anecdotal evidence of correct classification of policy positions for various legislators across the spectrum, ranging from the extremely left to the extreme right. By using Twitter based estimates, we see an improvement of classification of ideologies over the gold standard DW-NOMINATE score for legislators like Alexandria Ocazio-Cortez, whose are extreme left anti-moderate-Democrat votes seem to project closer to the Republicans. We also see that mean ideology has shifted slightly right over the years; though early adopters (2012) of political engagement on Twitter were dominated by liberals, we can see a shift towards the average non-partisan user being slightly Conservative. We realise that the defined model and the assumptions made about the model are reasonable by showing that the actual number of followers of a Twitter elite in the subset is not too far off from the value predicted by the model. We also show that standard errors provide valuable information about the

estimation of the ideal points - ideal points for ordinary users who are neutral/bipartisan are associated with higher variance as composed to users who have a stronger inclination to either ideology. The ability to analyse standard errors gives an advantage over other statistical, matrix-based and neural-network based methods. Although, methods from computer science using machine learning methods have an edge over this model, in predicting out-of-sample data, i.e., predicting ideology of an ordinary user given the information about which elites they follow. In my further work, I will try to incorporate/ extend the current Bayesian framework to be able to predict ideologies for out-of-sample data.

References

Bailey, Michael A. 2007. "Comparable Preference Estimates Across Time and Institutions for the Court, Congress, and Presidency." *American Journal of Political Science* 51 (3): 433–48. https://www.jstor.org/stable/4620077.

Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23 (1): 76–91. https://doi.org/10.1093/pan/mpu011.

Bonica, Adam. 2013. "Ideology and Interests in the Political Marketplace: *IDEOLOGY AND INTERESTS IN THE POLITICAL MARKETPLACE*." *American Journal of Political Science* 57 (2): 294–311. https://doi.org/10.1111/ajps.12014.

Desilver, Drew. 2015. "What Is the House Freedom Caucus, and Who's in It? Pew Research Center." October 20, 2015. https://www.pewresearch.org/fact-tank/2015/10/20/house-freedom-caucus-what-is-it-and-whos-in-it/.

Dilanian, Ken. 2019. "Is Trump's Pick for Top Spy Qualified for the Job? NBC News." July 29, 2019. https://www.nbcnews.com/politics/national-security/intel-officials-worry-trump-s-pick-top-spy-will-politicize-n1035821.

Gerrish, Sean M, and David M Blei. 2011. "Predicting Legislative Roll Calls from Text." In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 489–96. ICML'11. Madison, WI, USA: Omnipress.

GovTrack. 2020a. "Rep. Will Hurd [R-TX23]'s 2020 Legislative Statistics. — Govtrack.us." https://www.govtrack.us/congress/members/will_hurd/412654/report-card/2020.

———. 2020b. "Sen. Lisa Murkowski [R-AK]'s 2020 Legislative Statistics. — Govtrack.us." https://www.govtrack.us/congress/members/lisa_murkowski/300075/report-card/2020.

Hoff, Peter D, Adrian E Raftery, and Mark S Handcock. 2002. "Latent Space Approaches to Social Network Analysis." *Journal of the American Statistical Association* 97 (460): 1090–98. https://doi.org/10.1198/016214502388618906.

Homan, Matthew D., and Andrew Gelman. 2014. "The No-u-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *The Journal of Machine Learning Research* 15 (1): 1593–623.

Imai, Kosuke, James Lo, and Jonathan Olmsted. 2016. "Fast Estimation of Ideal Points with Massive Data." *American Political Science Review* 110 (4): 631–56. https://doi.org/10.1017/S000305541600037X.

———. 2022. *emIRT: EM Algorithms for Estimating Item Response Theory Models*. https://CRAN.R-project.org/package=emIRT.

Iyyer, Mohit, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. "Political Ideology Detection Using Recursive Neural Networks." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1113–22. Baltimore, Maryland: Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1105.

Kulkarni, Vivek, Junting Ye, Steven Skiena, and William Yang Wang. 2018. "Multi-View Models for Political Ideology Detection of News Articles." arXiv. https://doi.org/10.48550/arXiv.1809.03485.

Lahoti, Preethi, Kiran Garimella, and Aristides Gionis. 2018. "Joint Non-Negative Matrix Factorization for Learning Ideological Leaning on Twitter." In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 351–59. Marina Del Rey CA USA: ACM. https://doi.org/10.1145/3159652.3159669.

Lewis, Jeff. 2022. "Voteview | Why Is Alexandria Ocasio-Cortez Estimated to Be a Moderate by NOMINATE?"

January 20, 2022. https://voteview.com/articles/ocasio_cortez.

Lewis, Jeffrey B., Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet. 2023. "Voteview: Congressional Roll-Call Votes Database." Voteview. https://voteview.com.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *The Journal of Chemical Physics* 21 (6): 1087–92. https://doi.org/10.1063/1.1699114.

Poole, Keith T., and Howard Rosenthal. 1999. "Review of Congress: A Political-Economic History of Roll Call Voting." *Public Choice* 100 (1): 135–37. https://www.jstor.org/stable/30026084.

"ProgressivePunch: House Members by Score / All Issues. ProgressivePunch." 2023. 2023. https://progressivepunch.org/scores.htm?house=house.

Raju, Manu, and Darren Goode. 2011. "Murkowski Shows Independent Streak — Politico.com." https://www.politico.com/story/2011/05/murkowski-shows-independent-streak-055808.

Rauhauser, Neal. 2019a. "U.S. House Twitter followers March 2019," May. https://doi.org/10.6084/m9.figshare.8174672.v1.

———. 2019b. "U.S. Senate Twitter followers March 2019," May. https://doi.org/10.6084/m9.figshare.8170832.v1.

Siddique, Nauman. 2019. "2019-04-01: Creating a Data Set for 116th Congress Twitter Handles." April 2019. https://ws-dl.blogspot.com/2019/04/2019-04-01-creating-data-set-for-116th.html.

Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52 (3): 705–22. https://www.jstor.org/stable/25193842.

Wrubel, Laura, and Daniel Kerchner. 2020. "116th U.S. Congress Tweet Ids." Harvard Dataverse. https://doi.org/10.7910/DVN/MBOJNS.