Exploring New Ideal Point Estimation Strategies with Twitter Social Networks and Text Data

Yicheng Shen[1,3], Christopher A. Bail[2,3], and Alexander Volfovsky[1,3]
1. Department of Statistical Science, Duke University, Durham, NC, USA, 2. Department of Sociology, Duke University, Durham, NC, USA and 3. The Polarization Lab, Duke University, Durham, NC, USA

1. Introduction & Definition

- Analysis of network interactions on social media platforms, such as Twitter, Youtube and Facebook, has become a popular technique for studying the ideological orientation of netizens.
- An individual's **ideal point** is a summary of an individual's decision-making preferences within a spatial framework (Bafumi et al., 2005):
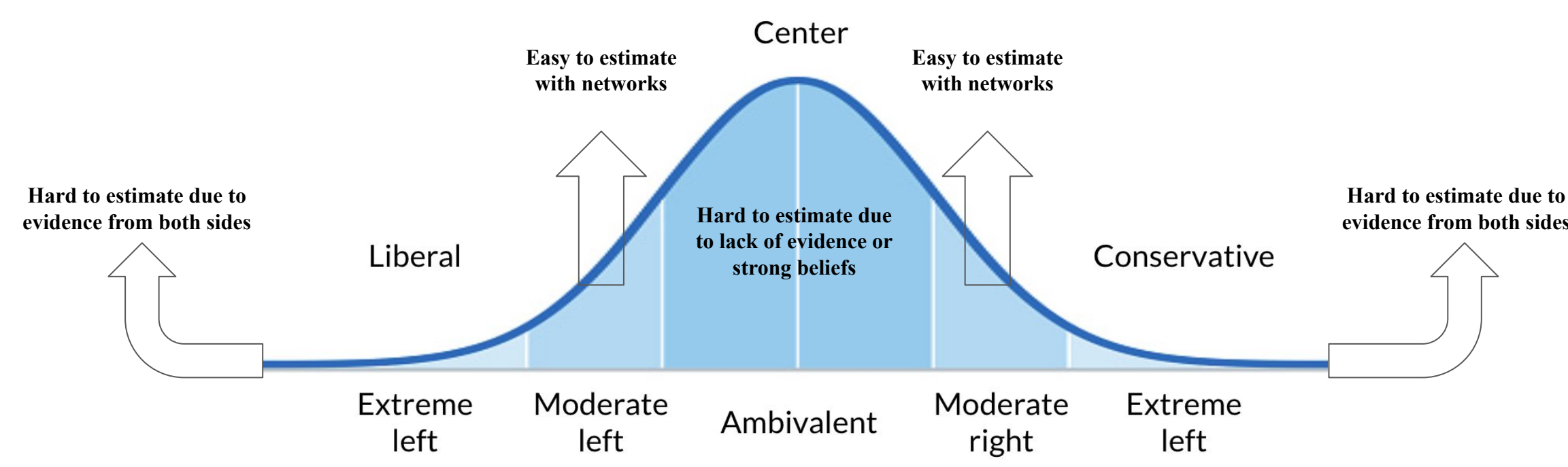
- Goal: estimate an individual's ideal point.
- Current state of the art based on social network data: a Bayesian spatial following model applied to network data (Barberá, 2015).
- Possible issues with ideal points that *current methods do not address*:
  - Do ideal points change over time?
  - Are extreme ideal points harder to estimate?
- We propose a newly refined approach that combines **network and text** data to improve the reliability of the estimates.

2. Literature & Motivation

- Many statistical methodologies have been developed to model ideal points in a multidimensional space, for example
  - *logistic regressions* (Wasserman and Pattison, 1996; Bafumi et al., 2005)
  - *covariate-based mixed-effects models* (Bailey, 2001)
  - *item response theory* (Clinton et al., 2004; Martin and Quinn, 2002)
  - *multinomial models* (Goplerud, 2019)
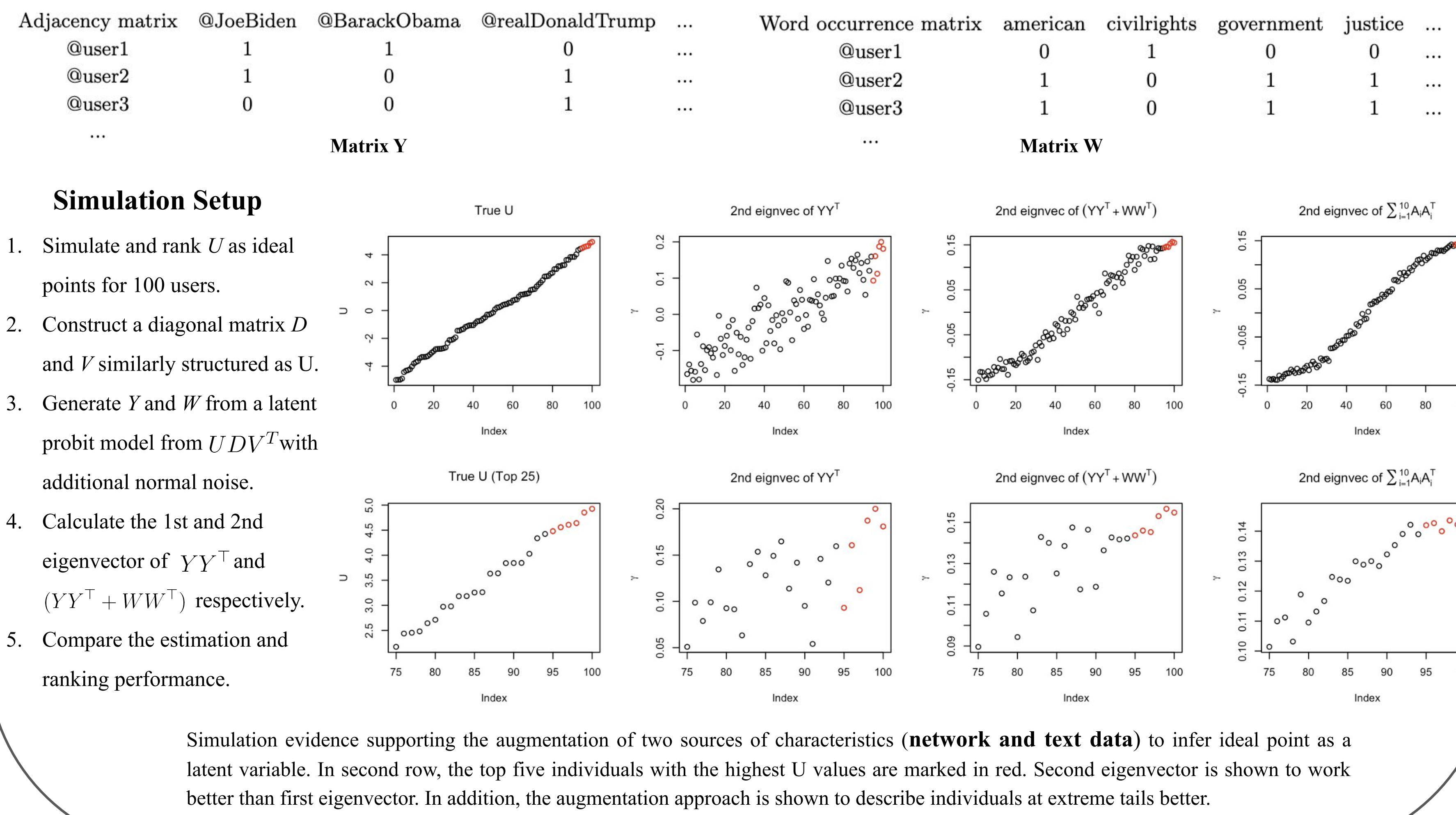 - *discriminant analysis* (Takane et al., 1987)
- Built upon Hoff et al. (2002), Barberá (2015) argues that ideology acts as a continuous latent variable from a multidimensional space. Proposes the following model:

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

- Because MCMC is computationally burdensome, it is suggested to use the correspondence analysis (CA) to estimate ideal points θ_i and ϕ_j. This simply requires calculating the singular value decomposition of a matrix of Y.
- The estimations results of CA and Bayesian approaches are **comparable**. But both methods are mostly applied under static settings.

3. Our method: Joint Modeling of Networks and Text

- Similar to the decision to follow elites' accounts, users' choice of words and language habits can also reflect ideological beliefs (Edelman, 1985; De Fina et al., 2011). We combine information from users' networks and additional, similarly structured information about users' posted texts in online discourses.

Adjacency matrix	@JoeBiden	@BarackObama	@realDonaldTrump	...
@user1	1	1	0	...
@user2	1	0	1	...
@user3	0	0	1	...
...				

Matrix Y

Word occurrence matrix	american	civilrights	government	justice	...
@user1	0	1	0	0	...
@user2	1	0	1	1	...
@user3	1	0	1	1	...
...					

Matrix W

Simulation Setup

1. Simulate and rank U as ideal points for 100 users.
2. Construct a diagonal matrix D and V similarly structured as U.
3. Generate Y and W from a latent probit model from UDV^T with additional normal noise.
4. Calculate the 1st and 2nd eigenvector of YY^T and $(YY^T + WW^T)$ respectively.
5. Compare the estimation and ranking performance.

Simulation evidence supporting the augmentation of two sources of characteristics (**network and text data**) to infer ideal point as a latent variable. In second row, the top five individuals with the highest U values are marked in red. Second eigenvector is shown to work better than first eigenvector. In addition, the augmentation approach is shown to describe individuals at extreme tails better.

4.1 Results - CA Method

- The estimated ideal points describe users' self-identified party affiliation and party strength well.
  - We verify a weak **backfire effect of exposure** to opposing political views from Bail et al. (2018).
  - The whole sampled population is estimated to experience a **liberal shift in ideological beliefs** from 2017 to 2022.

- Problems of CA method on ordinary users:
  - Estimations appear to be **inconsistent through time.**
  - No guarantee that the **latent political landscape remains constant**.
  - Difficult to estimate users who actively follow **diversified or few accounts.**
  - Moderates are harder to estimated with **information solely from networks.**
- Problems of CA method on elite accounts:
  - Observed shifts that are largely **not accounted for** by elites' true ideological dispositions.
  - Certain elites accounts (mostly Democrats) have substantially **larger follower counts, likely with mixed ideological backgrounds** recently.
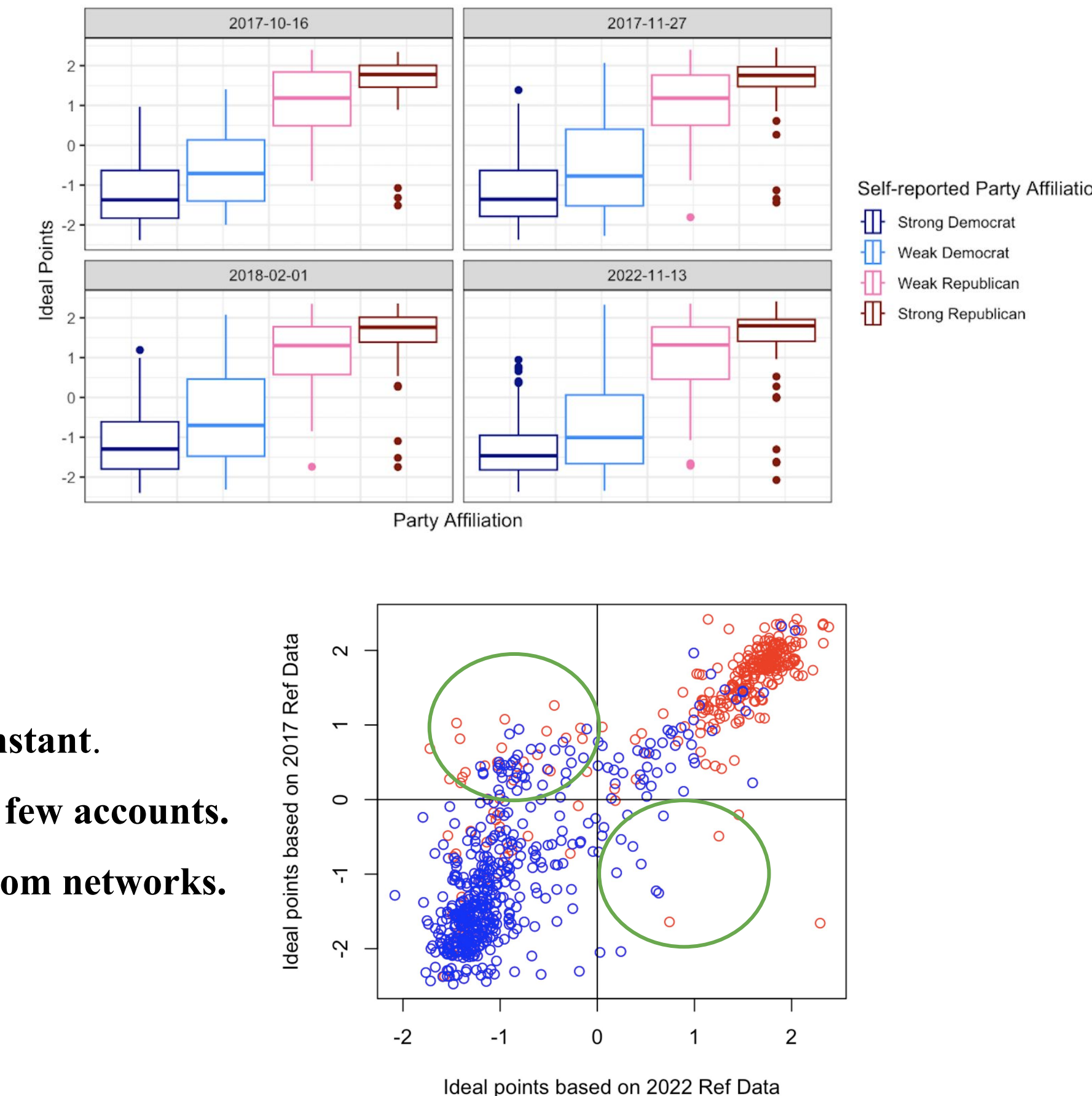

Twitter Account	$\hat{\gamma}$ in 2015	$\hat{\gamma}$ in 2022	Followers in 2015	Followers in 2022	Followers Increase Ratio
@JoeBiden	-1.2524	-0.6864	998,737	36,903,062	36.949729
@SpeakerPelosi	-1.1682	-0.6459	604,664	8,123,329	13.434451
@sensanders	-1.4355	-0.6767	349,149	12,494,541	35.785699
@HillaryClinton	-1.2127	-0.6078	3,612,817	31,580,745	8.741308
@FoxNews	0.7601	0.1514	5,235,703	23,711,845	4.528875
@TuckerCarlson	1.2155	1.6177	2,060,971	5,567,201	2.701252

Table 1. Selected elites' Twitter accounts with their ideal point estimations and number of followers taken based on network data collected in 2015 and 2022.
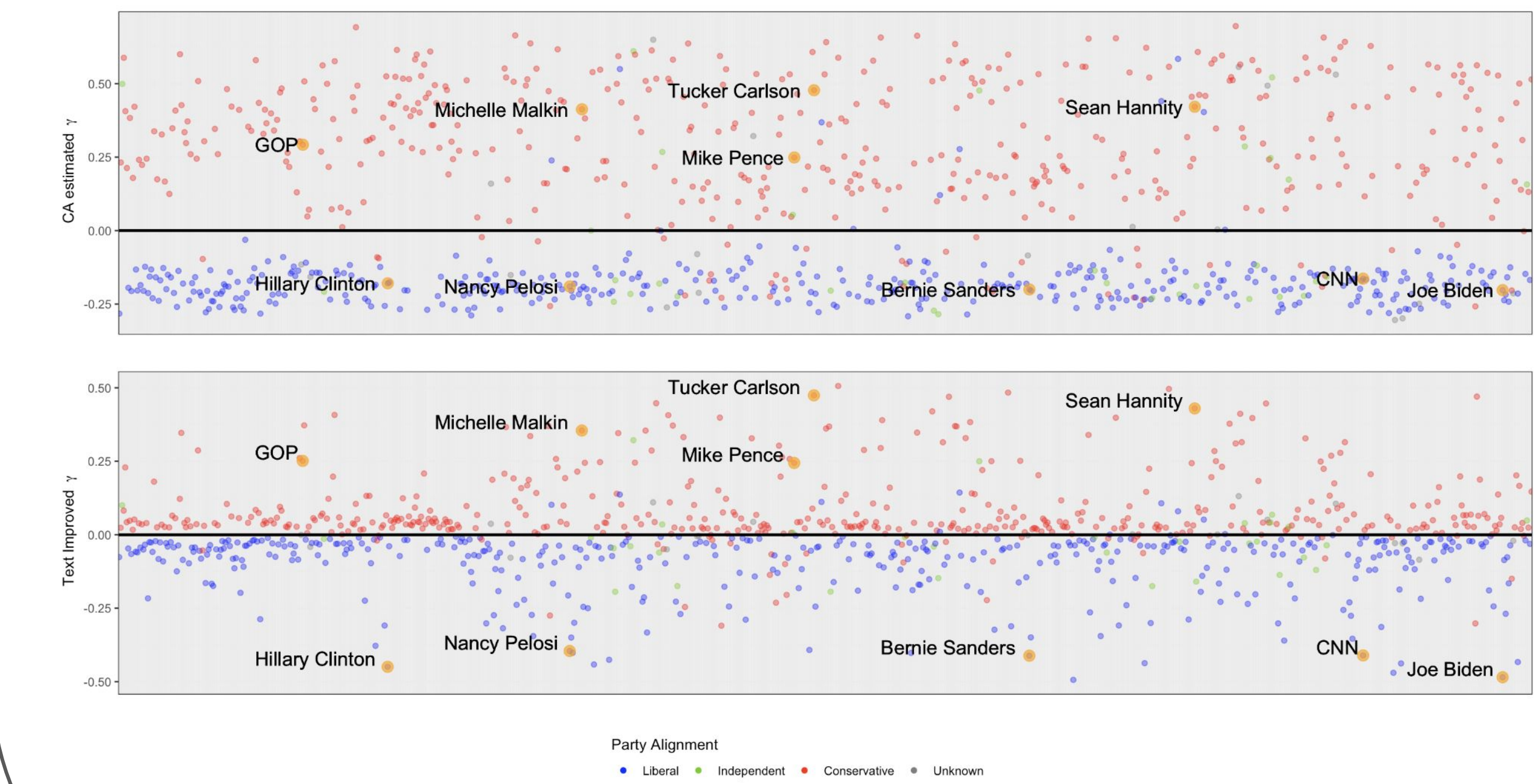
4.2 Results - Our method

The augmented method produces estimated ideal points that show:
- The majority of estimations **align with each other after** we apply text data.
- Elites on the far extreme ends of the ideological spectrum are estimated to be **more liberal** after considering their texts.
- Most moderate elites see a slight **conservative shift** with their posted texts.
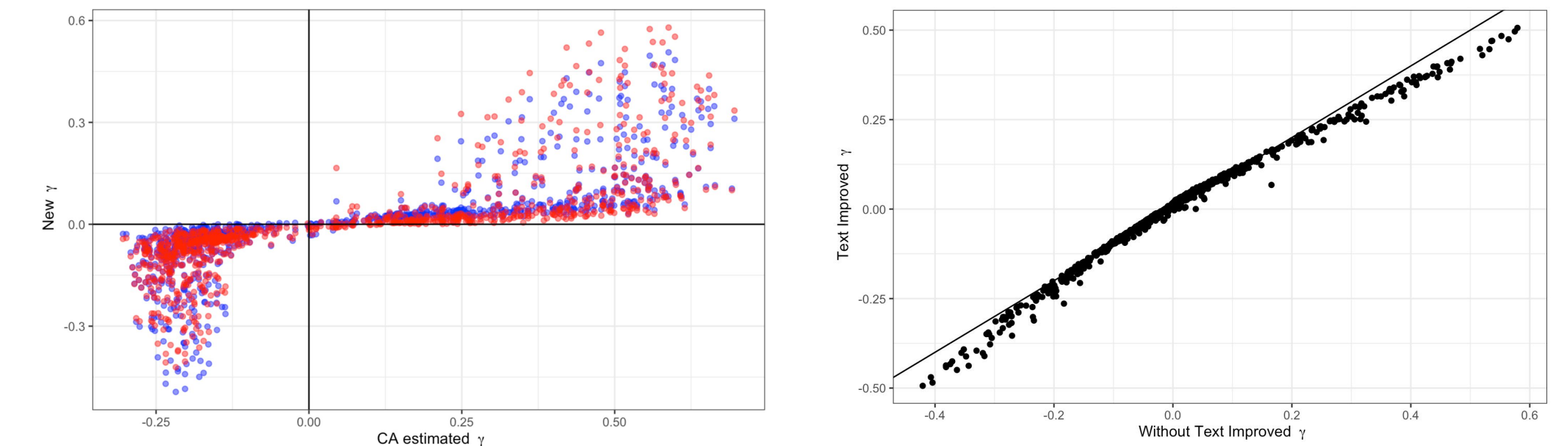
Our proposed method is able to
- Distinguish elites with **extremely liberal or conservative** ideologies better than CA method, especially political figures and media outlets that we can validate.
- Justify the augmentation of rich text data in the context of elite ideal point estimation by maintaining **few classification errors** verified by the true party alignments.

Elites with a large change of estimated ideal points after applying new method are highlighted in orange. Those with extreme ideologies are correctly adjusted and more distinct under the new method.

5. Conclusion

- Observe the **inconsistency of previously standard method** when applied on various time points with changing reference data.
- Integrate data from individuals' **Twitter networks and posted texts** to enhance ideal point estimation methodologies through a joint model.
- Successfully **highlight ideological extremity among political elites**, who were often estimated incorrectly by previous methods, and thus produce more reliable estimations to practitioners.

References
Bafumi, J., Gelman, A., Park, D. K., & Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, 13(2), 171-187.
Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., ... & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37), 9216-9221.
Bailey, M. (2001). Ideal point estimation with a small number of votes: A random-effects approach. *Political Analysis*, 9(3), 192-210.
Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political analysis*, 23(1), 76-91.
Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber?. *Psychological science*, 26(10), 1531-1542.
Clinton, J., Jackman, S., & Rivers, D. (2004). The statistical analysis of roll call data. *American Political Science Review*, 98(2), 355-370.
Goplerud, M. (2019). A Multinomial Framework for Ideal Point Estimation. *Political Analysis*, 27(1), 69-89.
Martin, A. D., & Quinn, K. M. (2002). Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953-1999. *Political analysis*, 10(2), 134-153.
Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460), 1090-1098.
Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3), 393-408.
Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p. *Psychometrika*, 61(3), 401-425.