Group 1 Project Proposal

Members: Sid Tekumalla, Sean George Dasovich, Lucas Sadoulet, Li Zhu

Prelude:

American politics is a frightful beast, but that hasn't stopped data scientists from wrangling it into a CSV. Our data comes from Voteview, a public database allowing users to view "every congressional roll call vote in American history on a map of the United States and on a liberal-conservative ideological map including information about the ideological positions of voting Senators and Representatives."

As indicated in our candidate dataset submission, our data is one specific dataset from this repository: the Member ideology of both the House and Senate from the 119th Congress (2025 - 2027) with the possibility of extending to previous Congresses.

Research Questions

1. Intra-Party Factions: Using the NOMINATE scores, can we identify distinct Republican factions (Freedom Caucus, moderates, Trump-aligned) and Democratic factions (progressives, moderates, Blue Dogs)? How accurate is this identification?
2. Age and Ideology: Using the birth year data, is there a correlation between member age and ideological positioning in 2025, controlling for party? Are younger Republicans/Democrats different from older ones, and in what ways?
3. Party Loyalty: Are politics getting more divided, i.e. is there less cooperation on bill / less breaking rank with each progressing congress?

Analysis Plan

We plan on answering the first research question using K-Means Cluster Analysis. Cluster analysis is an unsupervised machine learning approach that groups similar observations together based on their characteristics, without the use of predefined labels. The K-Means subtype involves the user specifying the number of clusters k (in this case, number of factions) and letting the algorithm assign each member to the nearest cluster center.

We plan on answering the second research question using Least Squares / Polynomial Regression. We don't know yet if the relationship between age and ideology is linear or polynomial. Meanwhile, making a scatterplot of age vs ideology scores will inform us how younger members think differently from their older peers.

We plan an answering the third using a partisanship metric. To determine how partisan an individual congress person is, we'll create a partisanship metric, based on how often the

person votes with their party, and the congress will receive a metric on how partisan it is on average.

To determine party affiliation or intra-party factions, we'll use both k-mean clustering and matrix decomposition to section out unknown variables and groupings.

On the topic of ideology…

Response to Teacher Feedback:

Q: I see a lot of different "ideology score" related variables in the dataset. How are they computed, and how are they different from [one] another? Which one do you plan to use as your response variable? Why?

There are two main score systems: NOMINATE and Nokken-Poole

- NOMINATE: a congressmember's average ideology in the entire Congress
    - Nominate_dim1 - measures liberal (-1) to conservative (+1)
    - Nominate_dim2 – historically measured regional (North/South) splits on issues, mainly civil rights. Now, however, what it measures is far more unclear. It's possible it represents establishment/moderates (+1) compared to populist/progressives (-1). Maybe a measure of "radicality"?
- Nokken-Poole: a congressmember's ideology at different points during Congress
    - Nokken_poole_dim1 – same as nominate_dim1, but allows score to change over time (captures ideological drifting)
    - Nokken_poole_dim2 – same as nominate_dim2, but is time-varying

We will use the NOMINATE score system for our purposes, likely favoring nominate_dim1 as it provides the most direct ideological score.