Supplementary S1:

Basic structural model

In STM, each speech can belong to a mixture of k topics. Topic proportions, θ_k, are allowed to correlate. Covariates are allowed to impact topical prevalence. The topical content covariate, party side in our case, allows term use within a topic to vary by content.

Let us assume that we have a corpus of A speeches, each of them indexed as $a_i \in (1, \ldots . A\ )$. Every parliamentary speech contains n observed terms. These are indexed as $n \in (1, \ldots, N_a)$. Each term, denoted by $v \in (1, \ldots, V\ )$, is part of the general vocabulary of the speech corpus. A term in the vocabulary is denoted by $w_{a,n}$. The number of topics, $k \in (1, \ldots \mathrm{K})$, is the most important input variable, which is drawn from a distribution. This distribution is affected by topic prevalence covariates (coalitions), which are specified in a $p \times 1$ vector X_a that contains party coalition information affecting the dominance of a topic k_i for each speech a_i.

We describe the process: First, the speech-level relation to each topic k is drawn from a logistic normal generalised linear model:

$$\vec{\theta}_a | X_{a\gamma} \sim \mathrm{LogisticNormal}_{K-1}\left(\mu X_{a\gamma}, \Sigma\right), \tag{1}$$

where γ is a $P \times (K-1)$ matrix of coefficients for the topic prevalence model drawn from a normal distribution for each k $(k = 1 \ldots K-1)$ with the other $K-1$ topics to provide bivariate dependence between topics. Σ is a $(K-1) \times (K-1)$ covariance matrix. In the next step, a topic-specific deviation from the first stage κ_k, a covariate for group deviation κ_g and an interaction term κ_i between them, using the speech-specific distribution over terms initially attributed to each topic (k) by the log frequency distribution (m) of the vocabulary vector, can be written as

$$\beta_{a,k,v} \propto \exp\left(m + \kappa_{k,v} + \kappa_{g,v} + \kappa_{i=k,g_a,v}\right), \tag{2}$$

where m, $\kappa_{k,v}$, $\kappa_{g,v}$ and κ_i are vectors of length V that contain one input per term in the vocabulary. Third, for each term $n \in (1, \ldots, N_a)$ in a speech a_i, the term specific topic assignment $z_{a,n}$ is modeled based on the speech-specific distribution over the given finite set of topics:

$$z_{a,n}|\vec{\theta}_\gamma \sim \text{Multinomial}_K(\vec{\theta}_\gamma), \tag{3}$$

The probability of an observed term to be attributed in this topic is then given by:

$$w_{a,n}|z_{a,n}, \beta_{a,k} = Z_{a,n} \sim \text{Multinomial}(\beta_{a,k} = Z_{a,n}). \tag{4}$$

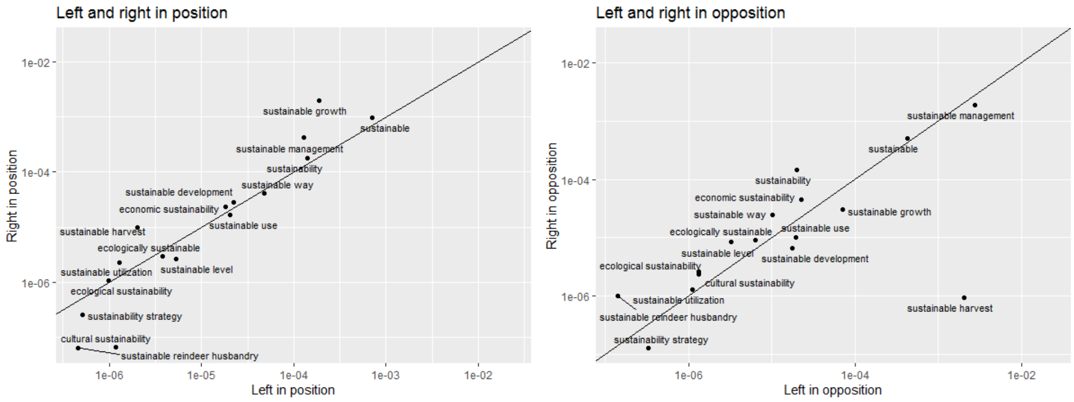
The model then is fit using the partially collapsed variational Expectation-Maximisation algorithm that uses a Laplace approximation to the non-conjugate portion of the model.

Supplementary S2

Figure S1. Scatterplots of log probabilities of sustainability terms by selected topic.

Petroleum

EU



Svalbard