# Modeling Ideological Salience and Framing in Polarized Online Groups with Graph Neural Networks and Structured Sparsity

**Valentin Hofmann**[*‡], **Xiaowen Dong**[†], **Janet B. Pierrehumbert**[†*], **Hinrich Schütze**[‡]

[*]Faculty of Linguistics, University of Oxford
[†]Department of Engineering Science, University of Oxford
[‡]Center for Information and Language Processing, LMU Munich
valentin.hofmann@ling-phil.ox.ac.uk

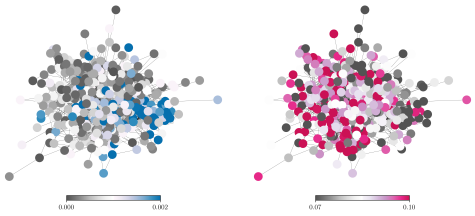
## Abstract

The increasing polarization of online political discourse calls for computational tools that automatically detect and monitor ideological divides in social media. We introduce a minimally supervised method that leverages the network structure of online discussion forums, specifically Reddit, to detect polarized concepts. We model polarization along the dimensions of salience and framing, drawing upon insights from moral psychology. Our architecture combines graph neural networks with structured sparsity learning and results in representations for concepts and subreddits that capture temporal ideological dynamics such as right-wing and left-wing radicalization.

## 1 Introduction

The polarization of online political discourse on platforms such as Twitter (Himelboim et al., 2013), Facebook (Bakshy et al., 2015), and Reddit (An et al., 2019) has received increasing attention in the computational social sciences recently, particularly in the context of Covid-19 (Green et al., 2020). In NLP, a growing body of work has discovered mechanisms by which polarization manifests itself linguistically (e.g., Demszky et al., 2019). However, the methods proposed so far rely on knowing in advance the political orientation of text, a requirement seldom met in social media.

In this paper, we propose SLAP4SLIP (Sparse LAnguage Properties for Social LInk Prediction), a novel framework that *fully dispenses with the need for labels* and instead leverages the ubiquitous network structure of online discussion forums to detect polarized concepts, making it more scalable and lightweight than previous methods. For example, SLAP4SLIP finds that *fascist* and *mainstream* are among the most polarized concepts in Reddit in 2019 (Figure 1). We model the polarization of concepts along the dimensions of salience and framing. For framing, we take into account insights about the moral foundations of ideology (Haidt and Joseph, 2004) and use contextualized BERT embeddings to construct subspaces that capture nuanced biases in the way concepts are discussed.



(a) *fascist* (salience) (b) *mainstream* (framing)

Figure 1: Examples of concepts polarized along the dimensions of salience (a) and framing (b) in Reddit in 2019. Each circle is a subreddit. The values for salience (a) are relative concept frequencies. References to fascism, reflected by higher relative frequencies of *fascist*, are typical for left-wing subreddits (blue region). The values for framing (b) are contextualized BERT embeddings projected into the moral sanctity/degradation subspace. The framing of *mainstream* as degenerate is pronounced in right-wing subreddits (magenta region). We can diagnose such patterns using SLAP4SLIP in a minimally supervised way.

**Contributions.** We introduce SLAP4SLIP, a framework to detect polarized concepts without information about the political orientation of text. The specific model we propose for SLAP4SLIP combines graph neural networks with structured sparsity learning and identifies in a minimally supervised way (i) which concepts are the most polarized ones, (ii) whether the polarization is due to differences in salience or framing, and (iii) which moral foundations are involved (when framing is relevant). Drawing on English Reddit data, we evaluate the model *intrinsically* by conducting various experiments and *extrinsically* by using the found polarized concepts to predict the ideological leaning of US states. The model also learns subreddit embeddings that capture temporal dynamics.[1]

[1]We make our code available at https://github.com/valentinhofmann/slap4slip.



*Findings of the Association for Computational Linguistics: NAACL 2022*, pages 536 - 550
July 10-15, 2022 

| Key term | Explanation |
|---|---|
| Polarization | By *polarization* we mean the clustering (of nodes, embeddings, etc.) according to ideology. Like Garcia et al. (2015), we understand it as a non-binary property, i.e., there can be more than two poles. A concept polarized in salience could be a unigram whose relative frequency has two clusters corresponding to liberalism and conservativism. |
| Salience | We understand *salience* as the topical prominence with which issues are discussed, indicating that a substantial importance is (consciously or unconsciously) ascribed to them. Issues that are highly salient (e.g., for an online group) tend to be mentioned often, which is reflected by word frequency statistics. |
| Framing | We use *framing* to refer to the mechanism by which certain aspects of an issue are highlighted. If framing patterns are exploited repeatedly (e.g., in an online group), this is reflected by word cooccurrence statistics. Due to the importance of moral foundations for ideological thinking, this paper focuses on moral framing. |

Table 1: Overview of our key technical terms. See main text for more details.

## 2 Related Work

Our study is closely related to previous NLP work on **polarization** (An et al., 2018; Demszky et al., 2019; Shen and Rosé, 2019; Roy and Goldwasser, 2020; Tyagi et al., 2020; Vorakitphan et al., 2020), but we try to avoid the need for explicit information about ideologies (e.g., manual labels) by leveraging the network structure of online discussion forums. Besides being more readily applicable in practice, this means our method is not restricted to a small number of opposing ideologies, making it theoretically more sound (Jackman, 2001). There is also work in the computational social sciences showing that the structure of various types of online social networks reflects polarization (Adamic and Glance, 2005; Garcia et al., 2015; Garimella et al., 2018), which has been explained as a result of homophily, i.e., nodes close to each other are likely to share similar views (McPherson et al., 2001). While these studies partition the network into a small number of ideological communities, our method does not require a discretization step. More broadly, our study is related to NLP work on **ideology** in general (Iyyer et al., 2014; Preotiuc-Pietro et al., 2017; Kulkarni et al., 2018).

Research in the political sciences has discovered **salience and framing** as two key dimensions along which the discussion of issues can vary ideologically. Salience refers to the amount of importance attached to an issue by individuals (Eulau, 1955; Miller et al., 2017). Mass media can impact salience, an effect called agenda setting (McCombs and Shaw, 1972). Framing refers to the mechanism by which certain aspects of an issue are highlighted (Entman, 1993; Druckman, 2001). Crucially, framing is different from sentiment: it reflects what considerations are perceived as important, not what stance is taken regarding these considerations (Nelson and Oxley, 1999). Both salience (with a focus on agenda setting) and framing have been the subject of previous work in NLP (Tsur et al., 2015; Card et al., 2016; Field et al., 2018; Mendelsohn et al., 2021). Here, we use them to characterize differences between online groups.

Psychological research has shown that the fundamental divisions between different ideologies are rooted in their views of morality (Lakoff, 2008). In **moral foundations theory** (Haidt and Joseph, 2004; Graham et al., 2011), this has been formalized as variation along the moral foundations of care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation. Several studies have shown that moral foundations theory is a suitable basis for analyzing ideological framing (Johnson and Goldwasser, 2018; Mokhberian et al., 2020; He et al., 2021). We follow this approach, but as opposed to prior work we operate with contextualized embeddings that we project into moral embedding subspaces.

Methodologically, we draw on advances in deep learning with **graph neural networks**, specifically graph auto-encoders (Kipf and Welling, 2016, 2017). In NLP, such graph-based architectures are increasingly used to include information from social networks for downstream tasks (e.g., Mishra et al., 2019; Hofmann et al., 2021). Our work differs in that we combine deep learning on graphs with **structured sparsity**, a form of regularization similar to $\ell_1$ regularization (Tibshirani, 1996) that sets entire groups of parameters to zero (Alvarez and Salzmann, 2016). Structured sparsity has been used in NLP before (Eisenstein et al., 2011; Murray and Chiang, 2015; Dodge et al., 2019), but not in connection with graph neural networks.

The precise definition of the key technical terms in this paper somewhat varies in the literature (e.g., Bramson et al., 2016). Table 1 therefore provides a short overview of how we use these terms.

## 3 SLAP4SLIP Framework

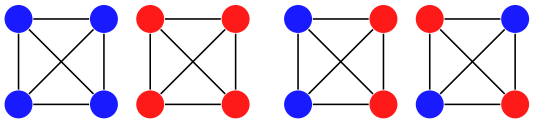
The key idea of this paper is to directly leverage the social network structure for determining polarized concepts.[2] We introduce a novel framework called SLAP4SLIP (Sparse LAnguage Properties for Social LInk Prediction) whose goal it is to model the structure of social networks in a data-driven way that obviates the need for extensive human annotation or partitioning the network into communities. SLAP4SLIP is a general framework to detect the most salient types of linguistic variablity in social networks and is in principle applicable in any scenario involving social networks with textual data attached to each node. In this paper, we show that for polarized online discussion forums, SLAP4SLIP can be used to find polarized concepts.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a network consisting of a set of nodes $\mathcal{V}$ representing social entities and a set of edges $\mathcal{E}$ representing connections between the social entities. We denote with $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ the adjacency matrix of $\mathcal{G}$. Let $\mathcal{C}$ be a set of word $n$-grams denoting concepts (e.g., political issues like *gun control*). Here, we confine ourselves to subreddits for $\mathcal{V}$ and unigrams and bigrams for $\mathcal{C}$, but SLAP4SLIP is applicable in other scenarios (e.g., for networks of people or concepts extracted from text in a more complex manner). We define a function $\psi_l : \mathcal{V} \times \mathcal{C} \rightarrow \mathbb{R}$ that assigns to each node $v_i \in \mathcal{V}$ and concept $c_j \in \mathcal{C}$ the value of a linguistic property $l$ observed for $c_j$ in $v_i$. $\psi_l$ can be represented as a matrix in $\mathbb{R}^{|\mathcal{V}|\times|\mathcal{C}|}$,

$$\mathbf{\Psi}_l = \begin{bmatrix} \psi_l(v_1, c_1) & \dots & \psi_l(v_1, c_{|\mathcal{C}|}) \\ \vdots & \ddots & \vdots \\ \psi_l(v_{|\mathcal{V}|}, c_1) & \dots & \psi_l(v_{|\mathcal{V}|}, c_{|\mathcal{C}|}) \end{bmatrix},$$

where each column is a graph signal (Dong et al., 2020) over $\mathcal{G}$ determined by $c_j$ and $\psi_l$. For example, if we chose $l$ to be the frequency count, $\psi_l$ would indicate how often each concept occurred in the text attached to each node of the network.

The goal of SLAP4SLIP is to find the subset of concepts $\mathcal{C}^* \subseteq \mathcal{C}$ that best meets the following two desiderata: (i) given a linguistic property $l$, the signals imposed on $\mathcal{G}$ by $\psi_l$ and the concepts in $\mathcal{C}^*$ should allow for optimal predictions about the structure of $\mathcal{G}$, specifically $\mathcal{E}$; (ii) the number of concepts in $\mathcal{C}^*$ should be minimal.[3] In practice, we treat this as a constrained optimization problem (Bertsekas, 1982), i.e., we use (i) as the objective and impose (ii) as a hard constraint on $|\mathcal{C}^*|$.



(a) $\psi_l$ for $c_1$ (b) $\psi_l$ for $c_2$

Figure 2: Example for the prediction of graph structure from a linguistic property. The figures show $\psi_l$ for concepts $c_1$ and $c_2$ on a toy graph, with $l$ chosen to be the frequency count represented by node color (identical colors mean identical frequencies). The edges can be fully predicted from $\psi_l$ for $c_1$ but not $c_2$.

As an example, consider the network in Figure 2. The network consists of two connected components of four edges each, with no edges between the components. $\mathcal{C}$ consists of the two concepts $c_1$ and $c_2$. Taking the frequency count as linguistic property $l$ and displaying it with the color of nodes, $\psi_l$ results in the two signals shown in Figure 2. We can see that the signal of concept $c_1$ alone allows for a perfect prediction of the network structure according to the decision rule

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if } \psi_l(v_i, c_1) = \psi_l(v_j, c_1) \\ 0 & \text{otherwise.} \end{cases}$$

Since $c_2$ cannot achieve a perfect prediction, $\mathcal{C}^* = \{c_1\}$ is the optimal solution. Notice the variance of $\psi_l(v_i, c_j)$ is identical for both concepts and does not represent a good distinguishing factor. Notice also that the optimal solution is not necessarily unique: there might be another concept $c_3$ with a similar frequency count distribution as $c_1$ such that $\mathcal{C}^* = \{c_3\}$ would also be an optimal solution.

## 4 Model

We draw upon Reddit Politosphere (Hofmann et al., 2022), a pseudonymized dataset based on Reddit covering 605 political subreddits (e.g., `politics`) from 2008 to 2019.[4] For each year, Reddit Politosphere contains (i) all comments made to the subreddits and (ii) an unweighted graph with the subreddits as nodes and edges computed by applying statistical backboning to the counts of users shared between subreddits. Subreddits that have disproportionately many users in common are likely to be ideologically similar (Kumar et al.,

[2] We define concepts as topics, issues, and public figures discussed in online groups.

[3] Desideratum (i) is conceptually similar to measures of opinion polarization on networks (Matakos et al., 2017).

[4] https://zenodo.org/record/5851729 (CC-BY 4.0 license)

2018). To ensure robust training, we only use years in which the graph has at least 100 nodes (2013 to 2019). See Appendix A.1 for summary statistics. The high modularity values indicate that the graphs are polarized (Kirkland, 2013).

We propose a neural architecture that uses information about concept-level salience and framing to predict links between subreddits while reducing the number of considered concepts as far as possible. Since the links reflect ideological similarity, this should result in a compact set of concepts that is maximally informative about ideology. The performance on link prediction makes it straightforward to compare the quality of different models.

**Determining concepts.** To obtain the concepts $\mathcal{C}$, we create for each year unigram and bigram vocabularies of political comments taken from Reddit Politosphere and non-political comments sampled in equal size from the default subreddits.[5] To eliminate unigrams and bigrams typical of discussions but not relevant to salience and framing (e.g., *dont think*), we only consider unigrams and bigrams that appear more often within than outside of noun phrases as detected by a noun phrase chunker (Honnibal et al., 2020). Based on their frequencies within the political and non-political comments, we compute mutual information scores for all unigrams and bigrams and take the top 1,000 unigrams and bigrams for $\mathcal{C}$. This and all other steps are done separately for each year, i.e., we extract year-wise concepts and train year-wise models.

**Modeling salience and framing.** The first part of the architecture models $\psi_l$, i.e., it extracts linguistic information related to salience and framing from the subreddits and maps them to scalar representations. In the resulting matrix $\mathbf{\Psi}_l$, each column is a signal on the entire graph defined by one concept, and each row is a vector for one subreddit defined by all concepts in $\mathcal{C}$ (Section 3).

To model ideological salience, we measure the relative frequency of concepts

$$s(v_i, c_j) = \frac{n(v_i, c_j)}{\sum_k n(v_i, c_k)},$$

where $n(v_i, c_j)$ is the frequency count of concept $c_j$ in subreddit $v_i$. Variations in the relative frequency of a concept that are strongly correlated with the social network structure indicate that the concept is used with systematically higher frequency in certain regions of the social network, potentially caused by its elevated place within the ideologies of the subreddits in question.

To model ideologically-driven framing, we use BERT (base, uncased; Devlin et al., 2019) and obtain average contextualized embeddings $\mathbf{e}(v_i, c_j)$ for each subreddit $v_i$ and concept $c_j$. Furthermore, we use the Moral Foundations Dictionary (Frimer et al., 2017) and obtain for each moral foundation $m_k$ (e.g., authority/subversion) average contextualized embeddings for the 10 highest-ranked words of both poles.[6] Similar to Bolukbasi et al. (2016), we perform PCA on the 20 average contextualized embeddings for each $m_k$ and use the first principal component as the subspace representation $\mathbf{e}(m_k)$. This allows us to project the subreddit-specific average contextualized concept embeddings $\mathbf{e}(v_i, c_j)$ into the five moral subspaces,

$$p_k(v_i, c_j) = \cos\left(\mathbf{e}(v_i, c_j), \mathbf{e}(m_k)\right).$$

$p_k(v_i, c_j)$ reflects how relevant the moral foundation $m_k$ is for the contexts in which concept $c_j$ occurs in subreddit $v_i$ (see Appendix A.2 for further details and a systematic evaluation). The moral foundations are expected to be relevant for the framing of concepts to differing degrees. We therefore compute concept-specific weighted sums,

$$f(v_i, c_j) = \sum_k \pi_k^{(c_j)} p_k(v_i, c_j),$$

where $\sum_k \pi_k^{(c_j)} = 1$ and $\pi_k^{(c_j)} \geq 0$. $f(v_i, c_j)$ is an aggregate indicator of how important moral framing is for concept $c_j$ in $v_i$. The parameters $\pi_k^{(c_j)}$ are optimized during training.

Salience and framing can be of different importance for different concepts, i.e., there might be concepts with identical values of $s(v_i, c_j)$ across all subreddits but maximally polarized values of $f(v_i, c_j)$ (or vice versa). To capture this, we combine $s(v_i, c_j)$ and $f(v_i, c_j)$ in a weighted sum,

$$o(v_i, c_j) = \alpha^{(c_j)} s(v_i, c_j) + (1 - \alpha^{(c_j)}) f(v_i, c_j),$$

where $0 \leq \alpha^{(c_j)} \leq 1$ is again a concept-specific parameter that is optimized during training. $o(v_i, c_j)$ indicates the overall activation of concept $c_j$ in $v_i$ (i.e., both due to salience and framing). Two important points must be stressed. First, $\pi_k^{(c_j)}$ and $\alpha^{(c_j)}$

[5] A set of topically diverse subreddits (e.g., `Fitness`) users used to be subscribed to automatically. We remove `news` and `worldnews` since they also contain political content. We retrieve the default subreddits from the Pushshift Reddit Dataset (Baumgartner et al., 2020).

[6] `https://osf.io/ezn37` (CC-BY 4.0 license)

are specific for concepts but identical for subreddits: e.g., if a concept $c_j$ has $\alpha^{(c_j)} = 1$, this means that only information from $s(v_i, c_j)$ is used for all subreddits. Second, values for $o(v_i, c_j)$ are comparable across subreddits but not across concepts: since $\pi_k^{(c_j)}$ and $\alpha^{(c_j)}$ differ between concepts, differences in $o(v_i, c_j)$ are not meaningful for different concepts (see Section 5 for examples). To get the final concept representation that is passed to subsequent parts of the model, we set $\psi_l = o$, i.e., each entry in $\mathbf{\Psi}_l$ contains the value of $o(v_i, c_j)$ for subreddit $v_i$ and concept $c_j$.

**Graph neural network.** To predict the links in $\mathcal{G}$, we use a graph neural network (Wu et al., 2021), specifically a graph auto-encoder (Kipf and Welling, 2016), which takes as input the matrix $\mathbf{\Psi}_l$ as well as $\mathcal{G}$'s adjacency matrix $\mathbf{A}$.

The encoder consists of a two-layer graph convolutional network (Kipf and Welling, 2017). In each layer, the subreddit representations $\mathbf{H}^{(d)}$ are updated according to the propagation rule

$$\mathbf{H}^{(d+1)} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(d)}\mathbf{W}^{(d)}\right),$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is $\mathcal{G}$'s adjacency matrix with added self-loops, $\tilde{\mathbf{D}}$ is the degree matrix of $\tilde{\mathbf{A}}$, and $\mathbf{W}^{(d)}$ is the weight matrix of layer $d$. $\sigma$ is the activation function, for which we use a rectified linear unit (Nair and Hinton, 2010) after the first and a linear activation (no non-linearity) after the second layer. We set $\mathbf{H}^{(0)} = \mathbf{\Psi}_l$. In our architecture, $\mathbf{Z} = \mathbf{H}^{(2)}$ is the output of the encoder. Graph convolutions are mathematically equivalent to Laplacian smoothing (Li et al., 2018), which is an important property for our architecture: if a concept does not occur in a subreddit, it ensures that the subreddit receives a high-quality representation by drawing on the neighboring subreddits.

In the decoder, we compute the reconstructed adjacency matrix, $\hat{\mathbf{A}}$, according to

$$\hat{\mathbf{A}} = \sigma\left(\mathbf{Z}\mathbf{Z}^\top\right),$$

where we use the sigmoid for $\sigma$. $\hat{\mathbf{A}}$ is then used to compute a prediction loss, $\mathcal{L}^{(\text{pred})}$.

**Structured sparsity.** Following the SLAP4SLIP framework, we want to reduce the number of concepts in $\mathcal{C}$. In the described architecture, this amounts to reducing the number of columns in $\mathbf{\Psi}_l$. We want to achieve this as part of training, using structured sparsity learning, specifically group lasso regularization (Yuan and Lin, 2006), to set entire rows of the weight matrix $\mathbf{W}^{(0)}$ to zero. Writing $\mathbf{W}^{(0)} = [\mathbf{w}_1^{(0)}, \ldots, \mathbf{w}_{|\mathcal{C}|}^{(0)}]^\top$ as a series of row vectors, we define the regularization penalty as

$$\mathcal{L}^{(\text{reg})} = \sum_{j=1}^{|\mathcal{C}|} \|\mathbf{w}_j^{(0)}\|_2.$$

This is a mixed $\ell_1/\ell_2$ regularization (the $\ell_1$ norm of the row $\ell_2$ norms) that leads to sparsity on the level of rows. When all entries in a row $\mathbf{w}_j^{(0)}$ are zero, this has the effect of removing concept $c_j$ from $\mathcal{C}$. We compute the final loss as

$$\mathcal{L}^{(\text{total})} = \mathcal{L}^{(\text{pred})} + \lambda\mathcal{L}^{(\text{reg})},$$

where $\lambda > 0$ is a hyperparameter controlling the intensity of the $\ell_1/\ell_2$ regularization.

## 5 Experiments

**Setup.** For each year, we split $\mathcal{E}$ into 60% train, 20% dev, and 20% test edges. We always use the train edges for the adjacency matrix $\mathbf{A}$ that is passed to the model, i.e., only the to-be-predicted edges differ between train, dev, and test. For dev and test, we randomly sample non-edges $(v_i, v_j) \notin \mathcal{E}$ as negative examples such that edges and non-edges are balanced in both sets (50% positive, 50% negative). For training, we sample non-edges in every epoch (i.e., the set of sampled non-edges changes in every epoch). During test, we rank all edges according to their predicted scores. See Appendix A.3 for hyperparameter details.

In this paper, we use sparsity as a hard constraint on the number of concepts with non-zero row weights in $\mathbf{W}^{(0)}$, i.e., we only consider models for which $|\mathcal{C}| \leq \theta_{|\mathcal{C}|}$, where $\theta_{|\mathcal{C}|}$ is the sparsity threshold. We initially set $\theta_{|\mathcal{C}|} = 150$ but later analyze its impact in greater detail.

The model is trained with binary cross-entropy as $\mathcal{L}^{(\text{pred})}$ and Adam (Kingma and Ba, 2015) as the optimizer. Since $\mathcal{L}^{(\text{reg})}$ is non-differentiable, we use proximal gradient descent (Parikh and Boyd, 2013). We approximate the weighted proximal operator of the $\ell_1/\ell_2$ norm using the Newton-Raphson algorithm (Deleu and Bengio, 2021). We use area under the curve (AUC) for model evaluation. We refer to our model as **SF-SGAE** (Salience/Framing Sparse Graph Auto-Encoder).

**Intrinsic evaluation.** We compare SF-SGAE against three ablated models: one where we use only salience, i.e., $\psi_l = s$ (S-SGAE), one where

| Model | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | $\mu \pm \sigma$ |
|---|---|---|---|---|---|---|---|---|
| SF-SGAE | .890 | .895 | .895 | .923 | .937 | .908 | .934 | .912±.018 |
| S-SGAE | .886 | .890 | .853 | .875 | .894 | .864 | .925 | .884±.022 |
| F-SGAE | .875 | .893 | .878 | .885 | .905 | .875 | .917 | .890±.015 |
| SF-SLAE | .653 | .810 | .754 | .781 | .764 | .729 | .752 | .749±.046 |
| SF-GAE | .829 | .797 | .871 | .916 | .898 | .866 | .933 | .873±.044 |

Table 2: Test performance (AUC). SF-SGAE outperforms S-SGAE, F-SGAE, and SF-SLAE. It performs similarly to or better than SF-GAE despite using only a fraction of concepts. Best score per column in gray. See Appendix A.4 for dev performance.

we use only framing, i.e., $\psi_l = f$ (F-SGAE), and one where we use both types of information but replace the graph convolutions with linear layers (SF-SLAE). Furthermore, we implement a model that is identical to SF-SGAE but does not use sparsity, i.e., $|\mathcal{C}|$ is not reduced (SF-GAE).

SF-SGAE clearly—and substantially on some years—outperforms the ablated models (Table 2). This shows that jointly modeling salience and framing captures polarization better than only modeling one of the two. Between S-SGAE and F-SGAE, there is no clear winner, although F-SGAE performs slightly better overall. SF-SLAE performs substantially worse than all other models, which indicates that the Laplacian smoothing in the form of graph convolutions is a crucial component of the model. SF-SGAE also outperforms SF-GAE on test, suggesting that $\mathcal{C}^*$ allows for a more robust generalization than the larger but noisier $\mathcal{C}$.

How does the sparsity threshold $\theta_{|\mathcal{C}|}$ impact model performance? The answer to this question indicates how many concepts are required to capture the central ideological divides in the data. We vary $0 \leq \theta_{|\mathcal{C}|} \leq 1000$ and measure the performance (AUC) of the four sparsifying models on dev (Figure 3). First, we find that for the models using graph convolutions, reducing $|\mathcal{C}|$ to approximately 200 concepts does not hurt performance. For the model without graph convolution, on the other hand, performance starts to drop already around 400 concepts. This makes intuitive sense: given that the graph convolutions act as a form of smoothing, less concepts are needed for a reliable feature vector for each subreddit. Second, the advantage of SF-SGAE lies not only in its higher performance in the sparse regime but also in its ability to reduce $|\mathcal{C}|$ much further than any of the other models given a performance threshold. This again demonstrates that a joint model of salience and framing results in richer information, making it possible to reduce the number of concepts further.

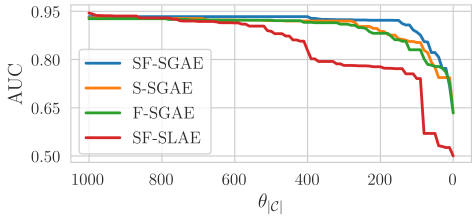



Figure 3: Impact of sparsity threshold $\theta_{|\mathcal{C}|}$ on performance (AUC) on dev for 2016. SF-SGAE performs better than any other model in the sparse regime ($\theta_{|\mathcal{C}|} \leq 200$), showing that it better captures polarization. Plots for all years are provided in Appendix A.5.

**Extrinsic evaluation.** The fact that SLAP4SLIP is a minimally supervised framework makes it challenging to evaluate the correctness of our model. While the performance on link prediction indicates how well $\mathcal{C}^*$ captures the polarized structure of the social network, it is not a direct measure of ideological polarization. There is also no ground-truth dataset against which $\mathcal{C}^*$ could be compared. We therefore devise an alternative extrinsic evaluation method. Specifically, we use DW-NOMINATE (Poole and Rosenthal, 1985, 1997), a quantitative measure of the ideological polarization of members of the US Congress based on their roll-call voting behavior. Recently, a large dataset of DW-NOMINATE scores has been made publicly available (Lewis et al., 2021).

We first create a dataset with all comments from subreddits dedicated to US state-level politics (e.g., `TexasPolitics`) in 2018.[7] We discard subreddits with less than 250 comments, resulting in a set of 28 subreddits. For each state, we then compute the average DW-NOMINATE score of its representatives in the lower house of the 116th US Congress (elected in November 2018). The average DW-NOMINATE is a continuous measure of the ideological leaning of a state and ranges between –0.399 for Massachusetts (very liberal) and 0.467 for Idaho (very conservative). Notice that this score reflects the state-level voting shares to a certain extent (since it is averaged over the representatives elected by a state) while at the same time being more fine-grained (since representatives of the same party can differ ideologically). Finally, for each state-level subreddit $v_i$, we extract $s(v_i, c_j)$ for (i) the $d$ concepts $c_j$ from $\mathcal{C}^*$ with the highest frequency across all state-level subreddits and (ii)

[7] We choose the larger subreddit in the case of multiple state-level subreddits. We retrieve the subreddits from the Pushshift Reddit Dataset (Baumgartner et al., 2020).

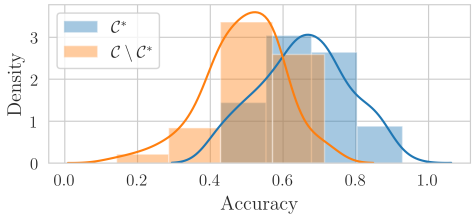



Figure 4: Performance on ideology prediction. The figure shows the distribution of accuracies for 100 models trained with relative frequencies of the concepts from $\mathcal{C}^*$ versus the concepts from $\mathcal{C} \setminus \mathcal{C}^*$. The concepts from $\mathcal{C}^*$ result in overall much higher accuracies, indicating that they better capture ideological polarization.

$d$ frequency-matched concepts $c_j$ sampled from $\mathcal{C} \setminus \mathcal{C}^*$.[8] We set $d = 5$.[9] If the concepts from $\mathcal{C}^*$ are better predictors of the average DW-NOMINATE scores than the concepts from $\mathcal{C} \setminus \mathcal{C}^*$, this indicates that the model has learned a correct split into more versus less polarized concepts.

To test this empirically, we compute the absolute value of Pearson's $r$ between $s(v_i, c_j)$ and the DW-NOMINATE scores. We find a higher correlation for the concepts from $\mathcal{C}^*$ ($\mu = 0.285$, $\sigma = 0.062$) than for the concepts from $\mathcal{C} \setminus \mathcal{C}^*$ ($\mu = 0.126$, $\sigma = 0.121$), a difference that is shown to be significant ($p < 0.05$) by a two-tailed $t$-test. This indicates that the concepts in $\mathcal{C}^*$ reflect the polarization of US politics better than the concepts in $\mathcal{C} \setminus \mathcal{C}^*$.

Furthermore, we try whether it is possible to predict the DW-NOMINATE scores from the relative concept frequencies. Specifically, we binarize the DW-NOMINATE scores by dividing them into the upper and lower half, thus resulting in a balanced dataset of more conservative and more liberal subreddits. We then train $\ell_2$-regularized logistic regression classifiers using the relative frequencies of the concepts from $\mathcal{C}^*$ and $\mathcal{C} \setminus \mathcal{C}^*$ as features. Since the dataset is small, we train 100 models on different random (label-stratified) splits of the subreddits into 50% training and 50% test. The models based on the concepts from $\mathcal{C}^*$ have substantially higher accuracies ($\mu = 0.657$, $\sigma = 0.122$) than the models based on the concepts from $\mathcal{C} \setminus \mathcal{C}^*$ ($\mu = 0.491$, $\sigma = 0.109$), a difference that is again shown to be significant ($p < 0.01$) by a two-tailed $t$-test (Figure 4). We interpret this as further evidence that the concepts in $\mathcal{C}^*$ (as opposed to the concepts in $\mathcal{C} \setminus \mathcal{C}^*$) capture ideological polarization.

[8] For $\mathcal{C}^*$, we only consider concepts for which $\alpha^{(c_j)} = 1$, i.e., the polarization is captured by $s(v_i, c_j)$ alone.

[9] Results are robust with respect to the exact selection of $d$.

| Year | $\alpha^{(c_j)} = 0$ | $0 < \alpha^{(c_j)} < 1$ | $\alpha^{(c_j)} = 1$ |
|---|---|---|---|
| 2013 | *aca* (l/b)<br>*bush* (a/s)<br>*tax* (c/h) | *deregulation* (l/b)<br>*fox news* (f/c)<br>*gun control* (l/b) | *gay marriage*<br>*gerrymandering*<br>*surveillance* |
| 2016 | *julian* (l/b)<br>*russian* (s/d)<br>*trump voters* (c/h) | *cuba* (a/s)<br>*gop* (s/d)<br>*nationalism* (l/b) | *collusion*<br>*fake news*<br>*reagan* |
| 2019 | *fact* (a/s)<br>*illegal* (a/s)<br>*mainstream* (s/d) | *congress* (a/s)<br>*white* (s/d)<br>*women* (c/h) | *donald*<br>*fascist*<br>*lefties* |

Table 3: Example concepts with $\alpha^{(c_j)}$ values of 1, 0, and in between. For $\alpha^{(c_j)} < 1$, we also provide the moral foundation $m_k$ with maximum $\pi_k^{(c_j)}$. c/h: care/harm; f/c: fairness/cheating; l/b: loyalty/betrayal; a/s: authority/subversion; s/d: sanctity/degradation. *aca* stands for Affordable Care Act (also known as Obamacare). *julian* refers to Julian Assange.

**Qualitative analysis.** We analyze which concepts are selected by SF-SGAE (Table 3). Many concepts in $\mathcal{C}^*$ are names of politicians (e.g., *bush*, *donald*) and designations of parties and political orientations (e.g., *gop*, *lefties*). Furthermore, $\mathcal{C}^*$ contains concepts related to contested political issues. While many of these issues (e.g., *gay marriage*, *gun control*) have been shown to be characterized by polarized online discussions before (Lai et al., 2015; Demszky et al., 2019), others (e.g., *deregulation*, *mainstream*) have been in the focus to a lesser degree, highlighting SLAP4SLIP's potential as an exploratory framework.

The design of our model also allows us to analyze in what way the concepts are polarized. To do so, we first examine the weight distribution of $\alpha^{(c_j)}$ for all $c_j \in \mathcal{C}^*$. We notice that for the majority of concepts (roughly 80%) $\alpha^{(c_j)} = 1$, i.e., the model uses only information about salience. Concepts with $\alpha^{(c_j)} = 1$ tend to be of immediate relevance for certain ideologies, leading to higher frequencies in relevant network regions. For example, in communist subreddits, discussion often revolves around fascism as the central opposing ideology, leading to higher frequencies of *fascist* than in other parts of the network (Figure 1a).

For concepts with $\alpha^{(c_j)} \neq 1$, we can analyze which moral foundation has the largest $\pi_k^{(c_j)}$. This moral foundation constitutes the basis for inter-subreddit differences in highlighting certain aspects of the concepts, which can be measured by $|p_k(v_i, c_j)|$, i.e., the absolute value of the projection of the concept embedding onto the $m_k$ subspace. For example, within the sanctity/degradation sub-

| | Small value of $\lvert p_k(v_i, c_j)\rvert$ | | Large value of $\lvert p_k(v_i, c_j)\rvert$ | |
|---|---|---|---|---|
| Concept $c_j$ | Subreddit $v_i$ | Example | Subreddit $v_i$ | Example |
| *bush* (2013, a/s) | `Freethought` | *This reminds me of what I read about the way the Bush administration worked religious quotes into military briefings.* | `Anarchy101` | *What's stopping from murderers becoming presidents? Oh wait... US has Obama, previously had Bush.* |
| *trump voters* (2016, c/h) | `Conservative` | *Trump voters, and people on the right in general, believe this is a grand country with little institutional racism left.* | `socialism` | *Trump voters have a hate boner for the Clintons that they've maintained since their 92 campaign.* |
| *mainstream* (2019, s/d) | `Kamala` | *She's good at making progressive ideas sound like reasonable mainstream policies, which is the best of both worlds.* | `TheNewRight` | *I think mainstream media has infected your brain with such rot that it effects your emotions.* |

Table 4: Polarization in framing. The table provides contexts for three concepts with $\alpha^{(c_j)} = 0$, both for subreddits with weak framing ($\lvert p_k(v_i, c_j)\rvert$ small) and subreddits with strong framing ($\lvert p_k(v_i, c_j)\rvert$ large) in the relevant moral subspace. c/h: care/harm; a/s: authority/subversion; s/d: sanctity/degradation.

space (the subspace with maximal $\pi_k^{(c_j)}$), many subreddits frame the concept *mainstream* in neutral terms. Right-wing subreddits, on the other hand, frame it as something degenerate, particularly in the context of media (Figure 1b, Table 4), reflecting appeals to discredit mainstream media reporting of political news (Lee and Hosam, 2020).

To get a more global picture of which moral subspaces are most important for the polarized framing, we examine the learned values of $\pi_k^{(c_j)}$ (Section 4) for all concepts with $\alpha^{(c_j)} \neq 1$. The three moral foundations that most frequently have the highest $\pi_k^{(c_j)}$ value are loyalty/betrayal (30%), sanctity/degradation (27%), and authority/subversion (21%), followed by care/harm (18%) and fairness/cheating (3%). Interestingly, loyalty/betrayal, sanctity/degradation, and authority/subversion are the three moral foundations with the greatest democrat-republican differences (Haidt and Graham, 2007; Graham et al., 2009), indicating that the US two-party system is a central axis for the polarized framing of concepts on Reddit.

**Ideological dynamics.** The embeddings $\mathbf{Z}$ learned by our model are subreddit representations that combine linguistic information with network information. Here, we analyze what types of temporal ideological dynamics are captured by $\mathbf{Z}$.

We map the embeddings $\mathbf{Z}$ for all years into a common embedding space using orthogonal Procrustes (Schönemann, 1966; Hamilton et al., 2016) and measure for each subreddit the cosine similarities between its embedding in the first year and its embeddings in all subsequent years. If the resulting time series of cosine similarities is continuously decreasing, this indicates a change in ideology. To detect such shifts automatically, we compute for each subreddit Pearson's $r$ between the time series of years and the time series of cosine similarities. Examining the subreddits with the most extreme negative values of $r$, we observe that most of them experienced a pronounced shift in their ideological orientation (Figure 5). Specifically, the subreddits move from a relatively moderate to a more extreme position in ideology space, either right-wing (e.g., `FreeSpeech`, `POLITIC`) or left-wing (e.g., `Sino`). This pattern suggests that the subreddits have ideologically radicalized over time (Grover and Mark, 2019; Youngblood, 2020).

## 6 Limitations

The success of our method depends on how accurately polarization is reflected by the network, which means that care must be taken during network selection (explicit networks) and construction (implicit networks). For example, user overlap on Reddit can also be due to conflict between subreddits (Datta et al., 2017; Kumar et al., 2018; Datta and Adar, 2019). While we do not find this to affect our results, it might be a limitation if the degree of homophily in the network is too low.

This paper only applies SLAP4SLIP to networks with communities as nodes and edges based on user overlap between the communities. However, the kind of clusteredness our method draws upon has been shown to be a property of various types of social networks, including social networks with individual users as nodes such as Twitter (Conover et al., 2011; Himelboim et al., 2013). We expect SLAP4SLIP to be a suitable framework for finding polarized concepts in these cases, too.

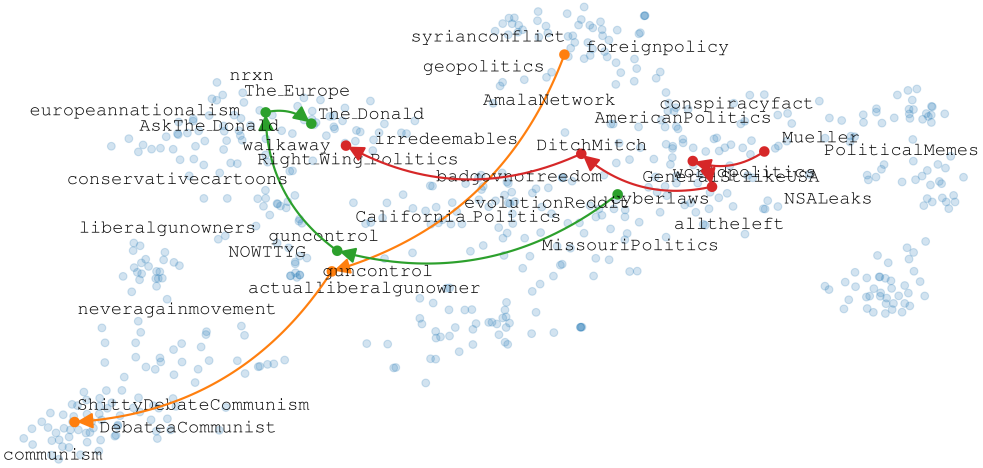



Figure 5: Example subreddits with a pronounced shift in their ideology over time. Orange: `Sino`, a subreddit originally devoted to geopolitics that moved to a more left-wing position; green and red: `FreeSpeech` and `POLITIC`, two originally moderate subreddits that moved to a more right-wing position.

## 7 Conclusion

We introduce SLAP4SLIP (Sparse LAnguage Properties for Social LInk Prediction), a novel framework for finding linguistic features maximally informative about the structure of a social network, and show that it can be used to detect polarized concepts. We model polarization along the dimensions of salience and framing. While we only address polarized concepts in this paper, the general nature of the framework makes it possible to apply it in diverse scenarios involving linguistic data attached to social networks (e.g., to find the most pronounced topical differences in citation networks). We see our study as an exciting first step towards bringing together computational social science research on online polarization, NLP work on political language, and graph-based deep learning.

## Ethical Considerations

As part of our model, we use contextualized word embeddings to model the polarized framing of concepts. However, contextualized word embeddings are known to be biased (Basta et al., 2019; Zhao et al., 2019; Bender et al., 2021), which bears the risk of impacting our results. We see this as an important research question for future work.

The user base of Reddit has been shown to be disproportionately young and male compared to the general population of the US (Shatz, 2017). We acknowledge that this limits the generalizability of our results, and we try to be particularly careful when drawing conclusions in the paper.

## Acknowledgements

This work was funded by the European Research Council (#740516) and the Engineering and Physical Sciences Research Council (EP/T023333/1). The first author was supported by the German Academic Scholarship Foundation and the Arts and Humanities Research Council. We thank the reviewers for their helpful comments. We are grateful to Jeremy Frimer for making the Moral Foundations Dictionary z-scores available to us.

## References

Lada A. Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 U.S. Election: Divided they blog. In *International Workshop on Link Discovery (LinkKDD) 3*.

Jose M. Alvarez and Mathieu Salzmann. 2016. Learning the number of neurons in deep networks. In *Advances in Neural Information Processing Systems (NIPS) 30*.

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Annual Meeting of the Association for Computational Linguistics (ACL) 56*.

Jisun An, Haewoon Kwak, Oliver Posegga, and Andreas Jungherr. 2019. Political discussions in homogeneous and cross-cutting communication spaces. In *International AAAI Conference on Web and Social Media (ICWSM) 13*.

Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and

opinion on Facebook. *Science*, 384(6239):1130–1132.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Workshop on Gender Bias in Natural Language Processing (GeBNLP) 1*.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift Reddit dataset. In *International AAAI Conference on Web and Social Media (ICWSM) 14*.

Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *ACM Conference on Fairness, Accountability, and Transparency (FAccT) 21*.

Dimitri P. Bertsekas. 1982. *Constrained optimization and Lagrange multiplier methods*. Academic Press, New York, NY.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems (NIPS) 30*.

Aaron Bramson, Patrick Grim, Daniel J. Singer, Steven Fisher, William Berger, Graham Sack, and Carissa Flocken. 2016. Disambiguation of social polarization concepts and measures. *The Journal of Mathematical Sociology*, 40(2):80–111.

Dallas Card, Justin H. Gross, Amber E. Boydstun, and Noah A. Smith. 2016. Analyzing framing through the casts of characters in the news. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2016*.

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Alessandro Flammini, and Filippo Menczer. 2011. Political polarization on Twitter. In *International AAAI Conference on Web and Social Media (ICWSM) 5*.

Srayan Datta and Eytan Adar. 2019. Extracting inter-community conflicts in reddit. In *International AAAI Conference on Web and Social Media (ICWSM) 13*.

Srayan Datta, Chanda Phelan, and Eytan Adar. 2017. Identifying misaligned inter-group links and communities. *Proceedings of the ACM on Human-Computer Interaction*, 1:1–23.

Tristan Deleu and Yoshua Bengio. 2021. Structured sparsity inducing adaptive optimizers for deep learning. In *arXiv 2102.03869*.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to Tweets on 21 mass shootings. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2019*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2019*.

Jesse Dodge, Roy Schwartz, Hao Peng, and Noah A. Smith. 2019. RNN architecture learning with sparse regularization. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2019*.

Xiaowen Dong, Dorina Thanou, Laura Toni, Michael Bronstein, and Pascal Frossard. 2020. Graph signal processing for machine learning: A review and new perspectives. *IEEE Signal Processing Magazine*, 37(6):117–127.

James N. Druckman. 2001. The implications of framing effects for citizen competence. *Political Behavior*, 23(2):225–256.

Jacob Eisenstein, Noah A. Smith, and Eric Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Annual Meeting of the Association for Computational Linguistics (ACL) 49*.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Heinz Eulau. 1955. Perceptions of class and party in voting behavior. *American Political Science Review*, 49(2):364–384.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: A computational analysis of intricate political strategies. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2018*.

Jeremy Frimer, Jonathan Haidt, Jesse Graham, Morteza Dehghani, and Reihane Boghrati. 2017. Moral foundations dictionaries for linguistic analyses, 2.0.

Dean Fulgoni, Jordan Carpenter, Lyle H. Ungar, and Daniel Preotiuc-Pietro. 2016. An empirical exploration of moral foundations theory in partisan news sources. In *International Conference on Language Resources and Evaluation (LREC) 10*.

David Garcia, Adiya Abisheva, Simon Schweighofer, Uwe Serdült, and Frank Schweitzer. 2015. Ideological and temporal components of network polarization in online political participatory media. *Policy and Internet*, 7(1):46–79.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27.

Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046.

Jesse Graham, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H. Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2):366–385.

Jon Green, Jared Edgerton, Daniel Naftel, Kelsey Shoub, and Skyler Cranmer. 2020. Elusive consensus: Polarization in elite communication on the COVID-19 pandemic. *Science Advances*, 6:eabc2717.

Ted Grover and Gloria Mark. 2019. Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. In *International AAAI Conference on Web and Social Media (ICWSM) 13*.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

William Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Annual Meeting of the Association for Computational Linguistics (ACL) 54*.

Zihao He, Negar Mokhberian, António Câmara, Andrés Abeliuk, and Kristina Lerman. 2021. Detecting polarized topics in COVID-19 news using partisanship-aware contextualized topic embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Itai Himelboim, Stephen McCreery, and Marc Smith. 2013. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication*, 18(2):40–60.

Valentin Hofmann, Janet B. Pierrehumbert, and Hinrich Schütze. 2021. Dynamic contextualized word embeddings. In *Annual Meeting of the Association for Computational Linguistics (ACL) 59*.

Valentin Hofmann, Hinrich Schütze, and Janet B. Pierrehumbert. 2022. The Reddit politosphere: A large-scale text and network resource of online political discourse. In *International AAAI Conference on Web and Social Media (ICWSM) 16*.

Matthew Honnibal, Ines Montani, Sofie van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Annual Meeting of the Association for Computational Linguistics (ACL) 52*.

Simon Jackman. 2001. Multidimensional analysis of roll call data via Bayesian simulation: Identification, estimation, inference, and model checking. *Political Analysis*, 9(3):227–241.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Annual Meeting of the Association for Computational Linguistics (ACL) 56*.

Diederik P. Kingma and Jimmy L. Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR) 3*.

Thomas N. Kipf and Max Welling. 2016. Variational graph auto-encoders. In *NIPS Bayesian Deep Learning Workshop*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR) 5*.

Justin H. Kirkland. 2013. Hypothesis testing for group structure in legislative networks. *State Politics & Policy Quarterly*, 13(2):225–243.

Vivek Kulkarni, Junting Ye, Steven Skiena, and William Y. Wang. 2018. Multi-view models for political ideology detection of news articles. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2018*.

Srijan Kumar, William Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *The Web Conference (WWW) 27*.

Mirko Lai, Cristina Bosco, Viviana Patti, and Daniela Virone. 2015. Debate on political reforms in Twitter: A hashtag-driven analysis of political polarization. In *International Conference on Data Science and Advanced Analytics (DSAA) 2015*.

George Lakoff. 2008. *The political mind: Why you can't understand 21st-century politics with an 18th-century brain*. Viking, New York, NY.

Taeku Lee and Christian Hosam. 2020. Fake news is real: The significance and sources of disbelief in mainstream media in Trump's America. *Sociological Forum*, 35:996–1018.

Jeffrey B. Lewis, Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet. 2021. Voteview: Congressional roll-call votes database.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *Conference on Artificial Intelligence (AAAI) 32*.

Antonis Matakos, Evimaria Terzi, and Panayiotis Tsaparas. 2017. Measuring and moderating opinion polarization in social networks. *Data Mining and Knowledge Discovery*, 31(5):1480–1505.

Maxwell E. McCombs and Donald L. Shaw. 1972. The agenda-setting function of mass media. *The Public Opinion Quarterly*, 36(2):176–187.

Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2021*.

Joanne M. Miller, Jon A. Krosnick, and Leandre R. Fabrigar. 2017. The origins of policy issue salience: Personal and national importance impact on behavioral, cognitive, and emotional issue engagement. In Jon A. Krosnick, I-Chant A. Chiang, and Tobias H. Stark, editors, *Political psychology: New explorations*, pages 125–171. Routledge, New York, NY.

Pushkar Mishra, Marco del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Abusive language detection with graph convolutional networks. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2019*.

Negar Mokhberian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. 2020. Moral framing and ideological bias of news. In *International Conference on Social Informatics (SocInfo) 12*.

Kenton Murray and David Chiang. 2015. Auto-sizing neural networks: With applications to $n$-gram language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2015*.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *International Conference on Machine Learning (ICML) 27*.

Thomas E. Nelson and Zoe M. Oxley. 1999. Issue framing effects on belief importance and opinion. *The Journal of Politics*, 61(4):1040–1067.

Mark Newman. 2018. *Networks*. Oxford University Press, Oxford, UK.

Neal Parikh and Stephen Boyd. 2013. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231.

Keith Poole and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American Journal of Political Science*, 29(2):357–384.

Keith Poole and Howard Rosenthal. 1997. *Congress: A political-economic history of roll call voting*. Oxford University Press, New York, NY.

Daniel Preotiuc-Pietro, Ye Liu, Daniel J. Hopkins, and Lyle H. Ungar. 2017. Beyond binary labels: Political ideology prediction of Twitter users. In *Annual Meeting of the Association for Computational Linguistics (ACL) 55*.

Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020*.

Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 36(1).

Itamar Shatz. 2017. Fast, free, and targeted: Reddit as a source for recruiting participants online. *Social Science Computer Review*, 35(4):537–549.

Qinlan Shen and Carolyn Rosé. 2019. The discourse of online content moderation: Investigating polarized user responses to changes in Reddit's quarantine policy. In *Workshop on Abusive Language Online 3*.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Annual Meeting of the Association for Computational Linguistics (ACL) 53*.

Aman Tyagi, Anjalie Field, Priyank Lathwal, Yulia Tsvetkov, and Kathleen M. Carley. 2020. A computational analysis of polarization on Indian and Pakistani social media. In *International Conference on Social Informatics (SocInfo) 12*.

Vorakit Vorakitphan, Marco Guerini, Elena Cabrio, and Serena Villata. 2020. Regrexit or not regrexit: Aspect-based sentiment analysis in polarized contexts. In *International Conference on Computational Linguistics (COLING) 28*.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2021. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24.

Mason Youngblood. 2020. Extremist ideology as a complex contagion: the spread of far-right radicalization in the United States between 2005 and 2017. *Humanities and Social Sciences Communications*, 7(1):310.

Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, 68(1):49–67.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2019*.

# A Appendix

## A.1 Data Statistics

Table 5 provides summary statistics of Reddit Politosphere (Hofmann et al., 2022). We compute average shortest path length as

$$\mu_\pi = \sum_{i,j \in \mathcal{V}} \frac{\pi(i,j)}{|\mathcal{V}|(|\mathcal{V}|-1)},$$

where $\pi(i,j)$ is the shortest path from subreddit $i$ to subreddit $j$. We compute density as

$$\rho = \frac{2|\mathcal{E}|}{|\mathcal{V}|(|\mathcal{V}|-1)}.$$

We compute modularity as

$$Q = \frac{1}{2|\mathcal{E}|} \sum_{i,j \in \mathcal{V}} \left( \mathbf{A}_{ij} - \frac{d_i d_j}{2|\mathcal{E}|} \right) \delta(i,j),$$

where $\delta(i,j) = 1$ if $i$ and $j$ are in the same community, else $\delta(i,j) = 0$. The maximum $Q$ values are indicative of the level of polarization in the graph. $Q > 0.3$ for all years, which is a typical cutoff value to determine polarized networks (Garcia et al., 2015). Notice we use the standard definitions of the three measures (Newman, 2018).

## A.2 Details on Moral Subspaces

For $\mathbf{e}(v_i, c_j)$, we extract the mean-pooled embedding if the concept is split into multiple WordPiece tokens and sample a maximum of 100 occurrences per subreddit and concept. For $\mathbf{e}(m_k)$, we sample 1,000 occurrences per word.

It is important to notice that $p_k(v_i, c_j)$ is impacted by two different factors. On the one hand, $p_k(v_i, c_j)$ captures the association of concepts with moral foundations due to *intrinsic* lexical-semantic properties, which can be seen by examining the variation of $p_k(v_i, c_j)$ across different concepts. Thus, computing $\frac{1}{|\mathcal{V}|} \sum_{v_i \in \mathcal{V}} p_k(v_i, c_j)$ for all concepts and moral foundations (i.e., the average value of $p_k(v_i, c_j)$ across subreddits), we find that the lexical semantics of concepts with the highest values are directly related to the moral foundations (e.g., *patriot* and *revolution* for loyalty/betrayal).

| Year | $\lvert\mathcal{D}\rvert$ | $\lvert\mathcal{V}\rvert$ | $\lvert\mathcal{E}\rvert$ | $\mu_d$ | $\mu_\pi$ | $\rho$ | $Q$ |
|---|---|---|---|---|---|---|---|
| 2013 | 6,306,458 | 108 | 324 | 6.00 | 3.08 | .056 | .560 |
| 2014 | 6,664,567 | 132 | 335 | 5.08 | 3.86 | .039 | .663 |
| 2015 | 9,230,022 | 168 | 493 | 5.87 | 3.87 | .035 | .672 |
| 2016 | 34,801,075 | 255 | 1,318 | 10.34 | 3.14 | .041 | .603 |
| 2017 | 38,278,685 | 295 | 1,572 | 10.66 | 3.14 | .036 | .585 |
| 2018 | 40,222,627 | 316 | 1,604 | 10.15 | 3.17 | .032 | .584 |
| 2019 | 46,590,000 | 412 | 2,536 | 12.31 | 3.20 | .030 | .603 |

Table 5: Dataset statistics. $|\mathcal{D}|$: number of comments; $|\mathcal{V}|$: number of nodes (subreddits); $|\mathcal{E}|$: number of edges; $\mu_d$: average node degree; $\mu_\pi$: average shortest path length; $\rho$: density; $Q$: maximum modularity.

On the other hand, $p_k(v_i, c_j)$ also captures the association of concepts with moral foundations that is due to *extrinsic* cooccurrence patterns caused by ideological framing, which can be seen by examining the variation of $p_k(v_i, c_j)$ across different contexts and subreddits (i.e., sets of contexts). To check this empirically, we use the 20 highest-ranked words per moral foundation from the Moral Foundations Dictionary (Frimer et al., 2017) and compute for each subreddit $v_i$, concept $c_j$, and moral foundation $m_k$ the proportion of occurrences in which at least one $m_k$ word is found in a context window of 10 words around $c_j$, which is similar to traditional ways of measuring ideological framing (e.g., Fulgoni et al., 2016). We then create for each concept $c_j$ and moral foundation $m_k$ (i) a set $\mathcal{T}_k(c_j)$ containing the $d$ subreddits with the largest proportion of moral context words and (ii) a set $\mathcal{B}_k(c_j)$ containing the $d$ subreddits with the smallest proportion of moral context words. We set $d = 5$, but results are robust with respect to the exact selection of $d$. Comparing the average value of $p_k(v_i, c_j)$ of subreddits in $\mathcal{T}_k(c_j)$ and $\mathcal{B}_k(c_j)$ for all concepts, we find it to be consistently higher for $\mathcal{T}_k(c_j)$ than for $\mathcal{B}_k(c_j)$ (Table 6). The fact that this result holds for all years and moral foundations suggests that the extent to which the concepts cooccur with certain moral frames is indeed captured by the projections of contextualized embeddings into the moral subspaces. Crucially, while $p_k(v_i, c_j)$ in principle captures both types of factors, only

| | Fairness/cheating | | | | | | | | Sanctity/degradation | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Set | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | $\mu \pm \sigma$ | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | $\mu \pm \sigma$ |
| $\mathcal{T}_k(c_j)$ | .074 | .074 | .074 | .075 | .076 | .076 | .076 | .075±.001 | .067 | .068 | .067 | .070 | .069 | .068 | .067 | .068±.001 |
| $\mathcal{B}_k(c_j)$ | .068 | .068 | .068 | .070 | .071 | .070 | .070 | .069±.001 | .065 | .064 | .064 | .067 | .066 | .065 | .064 | .065±.001 |

Table 6: Comparison of average $p_k(v_i, c_j)$ values for $\mathcal{T}_k(c_j)$ (large proportion of moral context words) and $\mathcal{B}_k(c_j)$ (small proportion of moral context words). The table shows the values for fairness/cheating and sanctity/degradation, but the trend is consistent across all moral foundations. Higher value per column in gray.

| Model | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | $\mu \pm \sigma$ |
|---|---|---|---|---|---|---|---|---|
| SF-SGAE | .857 | .893 | .911 | .921 | .923 | .913 | .921 | .906±.022 |
| S-SGAE | .833 | .868 | .872 | .864 | .883 | .865 | .904 | .870±.020 |
| F-SGAE | .832 | .880 | .863 | .861 | .884 | .868 | .894 | .869±.019 |
| SF-SLAE | .712 | .812 | .772 | .771 | .778 | .729 | .748 | .760±.031 |
| SF-GAE | .852 | .887 | .910 | .935 | .939 | .926 | .943 | .913±.031 |

Table 7: Dev performance (AUC). SF-SGAE outperforms S-SGAE, F-SGAE, and SF-SLAE. It performs similarly to or better than SF-GAE despite using only a fraction of concepts. Best score per column in gray.

the extrinsically-driven variation due to ideological framing is expected to be valuable for predicting the social network structure.

### A.3 Hyperparameters

The input layer of the model has 1,000 dimensions (which are sparsified during training), the first hidden layer 100 dimensions, and the second hidden layer 10 dimensions. We perform grid search for the number of epochs $e \in \{1, \ldots, 1000\}$, the learning rate $r \in \{1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}, 3 \times 10^{-3}\}$, and the regularization constant $\lambda \in \{1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}, 3 \times 10^{-3}\}$.

All experiments are performed on a GeForce GTX 1080 Ti GPU (11GB). The total number of trainable parameters is 107,110 for SF-SGAE, SF-SLAE, and SF-GAE, 101,110 for S-SGAE, and 106,110 for F-SGAE.

### A.4 Dev Performance

Table 7 provides the dev performance for all models considered in Section 5 of the paper.

### A.5 Sparsity Threshold

Figure 6 presents the results of the experiment varying the sparsity threshold described in Section 5 of the paper for all years.

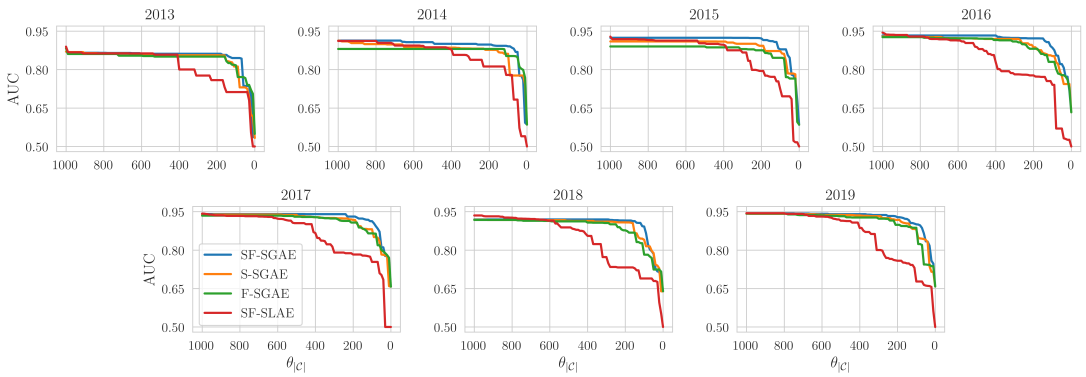



Figure 6: Impact of sparsity threshold $\theta_{|\mathcal{C}|}$ on performance (AUC). SF-SGAE performs better than any other model in the sparse regime ($\theta_{|\mathcal{C}|} \leq 200$), showing that it better captures polarization.