

State Preferences for Political Agendas and Policy Position Proximity: A Computational Content Analysis Project on the United Nations General Debate Corpus[★]

Luxin Tian, Heather Chen

MA Program in Computational Social Science

The University of Chicago

Abstract

Quantitative revelations of political agents' preferences are of substantial interest in the field of political science. In this project, we employ computational methods to perform a content analysis of the United Nations General Debate Corpus (UNGDC). We infer state preferences for political agendas through features embedded in text data and topic modeling, and we examine policy position proximity between states based on our measurement. We embed the documented statements delivered by countries into a semantic space and infer state blocs based on their policy positions represented by document vectors. We also present the topics that the states are concerned about over the past half-century and investigate international relations in terms of common concerns. Finally, we evaluate the validity and reliability of our methods and results.

Keywords: natural language processing, international relations, united nations

1. Introduction

There has been a growing interest in quantitatively measuring preferences of political agents in the field of political science. While qualitative approaches through case studies or historical analyses have been canonical in this field for centuries, quantitative measurement provides a more comprehensive and systematic framework in which the characteristics and dynamics revealed by actions of political agents can be identified, tracked, and analyzed. In the digital age, actions of political agents, such as parties, governments, and international organizations, can be traced in miscellaneous forms, and the development of data mining techniques and the computation capability unlocks new possibilities for researchers to reveal preferences of political agents through documented actions.

In this project, we perform an exploratory computational content analysis on the United Nations General Debate Corpus (UNGDC) to extract insights into state preferences for political

[★]This is a final project of the SOCI 40133 Computational Content Analysis course instructed by Prof. James Evans at the University of Chicago. Our teaching assistants are Bhargav Srinivasa Desikan and Hyunku Kwon.

agendas and positions from documented speeches. The UNGDC, which has been introduced and maintained by Baturo et al. (2017) and Jankin Mikhaylov et al. (2017), currently consists of 8093 statements made by state governments or heads of state at the annual General Debate of the United Nations General Assembly from 1970 to 2018. All the statements have been processed and converted to text files and organized by countries and sessions. Incorporating several prevailing techniques in Natural Language Processing, we make inference about international politics on two dimensions. First, we investigate the preferences for political agendas of international political agents, or, more specifically, the variation and dynamics of states' emphasis on certain topics of their concerns. Second, we extract position information of each state from the statements and quantify the dynamic similarities between states in political position space. As we will show in section 4, benefited from the inclusiveness and competitiveness of the UNGDC corpus, we can examine several aspects of contemporary international politics from a comprehensive perspective. Most of our data processing and computational analysis work is performed in Python.

As is argued by Baturo et al. (2017), the General Debate statements data has several favorable properties for comparative political analysis. For one thing, the lack of military allies for some states and the limited number of issues that are voted on in international conferences make such indicators incomplete and biased for measuring political preferences. In contrast, the General Debate provides all the member states of the United Nations an equal opportunity to address their positions on issues of their concerns to the international society. For another thing, due to loose institutional connection to the decision-making process, the General Debate statements are less constrained by external pressures from other states than voting actions are so that the contents would reflect more about the real interests and positions of states. These features make it more reliable and complete to use the UNGDC for revealing and measuring political preferences than using military alliances or voting records as proxy indicators.

In light of the advantages of the UNGDC, we assume through the whole analysis that the UNGDC is complete and unbiased for revealing political preferences and ideologies in a way that it contains semantically meaningful sentences that address political positions and concerns. This assumption may not be completely true. First, as we will show, the lengths for statements become more and more limited as the number of participating states increases over the years. As a result, since states may have priorities on elaborating certain issues over others, the compressed content may not represent the complete ideologies on all issues, and the distribution of emphasis on each topic can also vary across states and sessions. Second, the states may still face external pressures from other states when delivering the statements even though the opportunity is institutionally equal for all member states. This can result that the semantic meaning of one state's expressions is not independent but correlated to other states. If it is the case, the robustness and reliability of our inference would be qualified. However, we argue that the UNGDC still outperforms many other data sources for empirical analysis of state political preferences for the reasons mentioned above, and we will address potential drawbacks and our reflections of this research in sections 5 of this report.

In section 3, we review the computational methods that we employed in this project and evaluate the fitness of each method to our research question. For some of the methods, we also introduce our implementation of the algorithms and evaluate the computational performance of the program. In section 4, we present our findings and relate some of the results to historical events and con-

texts. Finally, in section 5, we assess the reliability and validity of our analysis, clarify potential drawbacks, and propose future extensions of our research.

2. Data

The United Nations General Debate Corpus (UNGDC) has been published and maintained by Jankin Mikhaylov et al. (2017). The latest update of the corpus consists of 8093 statements delivered by 200 states on the annual sessions of the United Nations General Assembly from 1970 to 2018. The comprehensive historical miscellaneous data for each session, based on which this corpus was built, is published on the website of the General Assembly of the United Nations in the forms of full-text .pdf documents, audio, video, and summary text in all the six official working languages. The UNGDC incorporates all the English version documents that record the statements made by state leaders or government representatives. Each document has been converted to a .txt file and categorized and labeled by an alpha-3 country code (ISO 3166), the session number, and the year.

The number of states that deliver the statements on each session varies across years, but overall it has been gradually increasing from 71 in 1970 to 200 in 2018. We identify these changes and include a table showing the change of participating states for each year in the appendix.

We preprocess the corpus by tokenizing the words in each document into lists of tokens, normalizing the tokens, removing digits, punctuations, non-Latin characters, and English stop words, and then lemmatizing. For some of the modeling procedures that we introduce in the next section, we also perform stemming on the normalized tokens, which reduces the tokens into their base forms. For tokenization and normalization, we take advantage of the helper functions from the `lucem_illud_2020` package. For stemming we use the PorterStemmer implemented by the NLTK package.

After tokenizing, we count the number of tokens in each document and calculate the mean token count per statement. As is shown by figure 1, with the number of participating states increasing over years, the mean token count steadily reduced almost in half, indicating the speakers face more and more rigorous time limitation when delivering the statement speech. However, the mean token count remains as high as around 2,000.

We arrange this corpus and organize it into the following format. For each document, we label it with a three-letter country code, an integer indicating the session, and another integer indicating the year. We save the results of tokenization and normalization into two separate columns for future use. Note that in the original UNGDC corpus one text document `URY_40_1985.txt` was mistakenly named as `URY_49_1985.txt`. We have corrected it by hand.

3. Methods and Implementations

The UNGDC corpus contains rich information about national political positions on a wide range of issues and topics. In this section, we briefly review the methods we employ to gain insights into the social mechanism underlying the data generating process of the text. First, we examine word frequency. On this basis, we identify the most common words conditional on their functions as parts of speech to reveal the most significant object being discussed in recent decades.

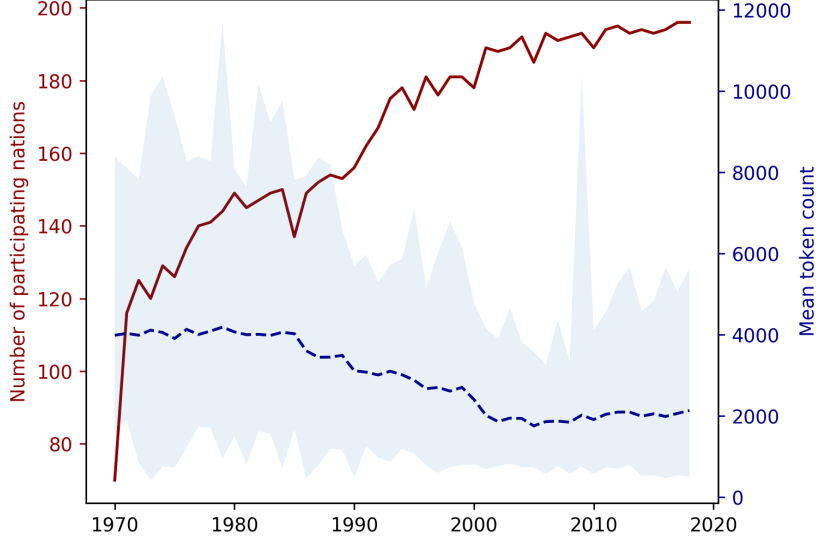


Figure 1: Number of Participating Nations and Mean Token Count per Statement (1970-2018)

We also try to directly quantify the proximity of state positions by extracting the principal components of the word frequency vectors as well as performing a K-means clustering on the vector space. Second, we embed the tokens onto a semantic space using the techniques named Word2Vec and Doc2Vec, based on which we investigate position proximity between states from a semantic perspective. Third, we turn to topic modeling and employ the Latent Dirichlet Allocation model to figure out the topics that the states concerned across years. Finally, we also construct international networks based on mutual information revealed by topic modeling to measure position proximity in terms of topics of common interests.

3.1. Word Frequency: Term Frequency - Inverse Document Frequency (TF-IDF)

The Term Frequency - Inverse Document Frequency (TF-IDF) measures the frequency of a word in a document normalized by the importance of documents in the whole corpus. Intuitively, it reduces the weights of the frequency of the words in a document that occur very frequently in the whole corpus. To characterize this normalized measurement of word frequency, for each token t in a document d , the formula of TF-IDF is given by:

$$\begin{aligned}
 \text{TF-IDF}(t, d) &= \text{Term Frequency (TF)} \times \text{Inverse Document Frequency (IDF)} \\
 &= \frac{\text{count of } t \text{ in } d}{\text{number of tokens in } d} \times \ln\left(\frac{\text{count of } t \text{ in corpus} + 1}{N}\right)
 \end{aligned} \tag{1}$$

We use `sklearn.feature_extraction.text.tfidfvectorizer` to extract the word frequency features for each token in each document of our corpus, which gives a document-token

Table 1: Preview of the UNGDC corpus

| filename | country_code | session | year | text |
|-----------------|--------------|---------|------|---|
| ALB_25_1970.txt | ALB | 25 | 1970 | 33: May I first convey to our President the co... |
| ARG_25_1970.txt | ARG | 25 | 1970 | 177.\t : It is a fortunate coincidence that pr... |
| AUS_25_1970.txt | AUS | 25 | 1970 | 100.\t It is a pleasure for me to extend to y... |
| AUT_25_1970.txt | AUT | 25 | 1970 | 155.\t May I begin by expressing to Ambassado... |
| BEL_25_1970.txt | BEL | 25 | 1970 | 176. No doubt each of us, before coming up to ... |
| BLR_25_1970.txt | BLR | 25 | 1970 | \n71.\t. We are today mourning the untimely de... |
| BOL_25_1970.txt | BOL | 25 | 1970 | 135.\t I wish to congratulate the President o... |
| BRA_25_1970.txt | BRA | 25 | 1970 | 1.\tMr. President, I should like, first of all... |
| CAN_25_1970.txt | CAN | 25 | 1970 | \nThe General Assembly is fortunate indeed to ... |
| CMR_25_1970.txt | CMR | 25 | 1970 | : A year ago I came here as the Acting Preside... |

TF-IDF matrix. Each row of this matrix can be viewed as a feature vector characterizing each document in terms of the relative significance of its tokens.

We assume that the content of a statement can be represented by its TF-IDF vector, which is expected to contain multidimensional information about the topics that the state concerns and their position and ideology on the topics. Even though the topics are not directly observable nor identifiable by the TF-IDF vector, we can compare any two documents by comparing the two TF-IDF vectors to infer the similarity of the content in the normalized word frequency space.

Based on the TF-IDF matrix, we perform two analyses. First, we reduce the dimensionality of the vectors using Principle Component Analysis (PCA) to unidimensional scalars. We expect this to abstractly represent the content of the statements in the TF-IDF space, which is equivalent to project multi-dimensional state preference to one single visible dimension. Then, we take several countries as examples and visualize the revelation of several countries' preferences across years. Second, we perform K-means clustering using the TF-IDF vectors associated with each document. We review and evaluate the K-means Clustering method in section 3.4

3.2. Part-of-Speech Tagging

Part-of-Speech (POS) Tagging is to parse a document of natural language and tag the grammatic role of each token in a sentence. Since it depends on the grammatic context to disambiguate the word category, we use the SpaCy's implementation of POS tagging to parse our documents using the original text without lemmatization or normalization. Note that one of the most frequent words "cooperation" is commonly written as "co-operation" in our corpus, and the POS tagger fails to identify it as a single word, we have corrected it to "cooperation" by hand before performing POS tagging. Conditional on word categories, we can count token frequency and figure out the most frequently-used word of certain grammatic category in our corpus, which can reflect the general topics and the main content of this international politics corpus.

3.3. Word Embeddings and Semantic Space: Word2Vec and Doc2Vec

Word embedding is a language modeling technique that maps words to vectors in a semantic space, of which Word2Vec and Doc2Vec are two computational implementations. In the resulting

vector representations of words, similar words with similar semantic meanings turn out to be close to each other (Mikolov et al., 2013). While Word2Vec embeds each token into a semantic space and represents tokens as numeric feature vectors, Doc2Vec enables us to label each document as a set of tokens and create numeric representations of documents in the same semantic space where the words are embedded. As a result, the word vectors can be thought of as representations of the semantic meaning of every single word, and the document vectors represent the meaning of each document.

In our analysis, we train a Doc2Vec model with our corpus using the implementation by `gensim.models.doc2vec`. Based on the resulting semantic space, we can quantify the proximity of any two states' position by calculating the cosine similarity of their document vectors. We can also assign hierarchical labels to the documents to integrate the features of all the documents of one state across years into a single vector. On this basis, we can project the states onto certain dimensions we define to answer our research question, for example, the Cold War dimension specified by the differences between the vectors of the NATO and the WTO states as well as 'capitalism' and 'socialism'. Furthermore, we perform K-means clustering using the document vectors to identify state blocs in terms of their overall standpoints of the Cold War extensive issues.

To verify that our corpus is complete and rich enough to construct a semantic space with meaningful results, we pre-train a Word2Vec model and experiment with several vector operations. We can find the most semantically similar words to a word. We can also perform vector operations to find the most matching word to an equation such as king – queen = man – women, which is one of the examples presented by Mikolov et al. (2013). We have many interesting results (shown in Table 2 and 3).

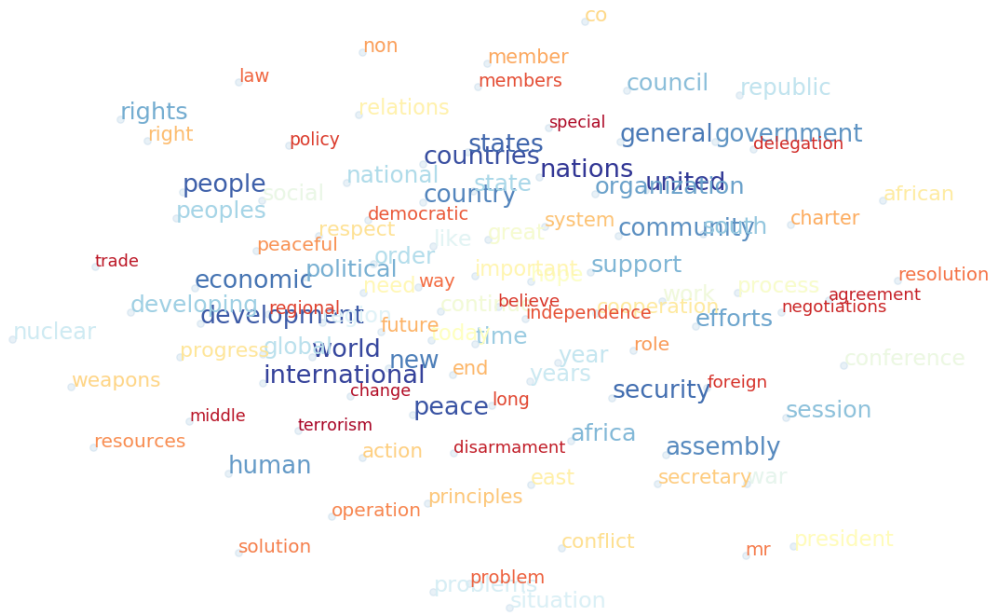
Table 2: Examples of finding the most similar words

| target word | similar words |
|-------------|--|
| economic | socioeconomic, macroeconomic, economy, economies |
| political | institutional, geopolitical, politico, civic |
| socialism | communism, revolution, militarism, fascism |

Table 3: Examples of vector operation in the semantic space

| token1 | – | token2 | = | token3 | – | token4 |
|-----------|---|----------|---|---------|---|----------|
| american | - | america | = | korean | - | korea |
| islamabad | - | pakistan | = | bangkok | - | thailand |
| oppose | - | support | = | reject | - | welcome |

We can also visualize the semantic space where the words or documents are embedded. Since the vectors are high dimensional, we perform dimension reduction using the Principle Component Analysis (PCA) implemented by `sklearn.decomposition.PCA` and the t-distributed Stochastic Neighbor Embedding (tSNE) implemented by `sklearn.manifold.TSNE`. While PCA is a technique for linear dimensionality reduction, tSNE is a specific tool to visualize high-dimensional



$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|} \in [0, 1] \quad (2)$$

, where \mathbf{a} and \mathbf{b} are any two vectors in the corpus. The higher the cosine similarity, the more similar the two vectors are in terms of semantic meanings.

We perform two analyses. First, we examine the similarities between states across years by calculating pairwise cosine similarities of the state vectors. We visualize the result in a heatmap. To alleviate noisy variations across years, we also show the average cosine similarities between states every five years to detect the most significant dispute and consensus. We will give the standard deviation of the cosine similarities at the end of the analysis, which indicates the variation of international relations in terms of policy proximity.

Second, we investigate the dynamics of international attention paid on specific policy areas by calculating the cosine similarities between the year vectors and the vectors representing the keywords of several policy topics. For the topics characterized by more than one token, we use the mean of multiple vectors to construct the basis vector. We can also observe the attention to such topics paid by specific states and evaluate to what extent these states are leading the international discussion of the policy issues in the United Nations.

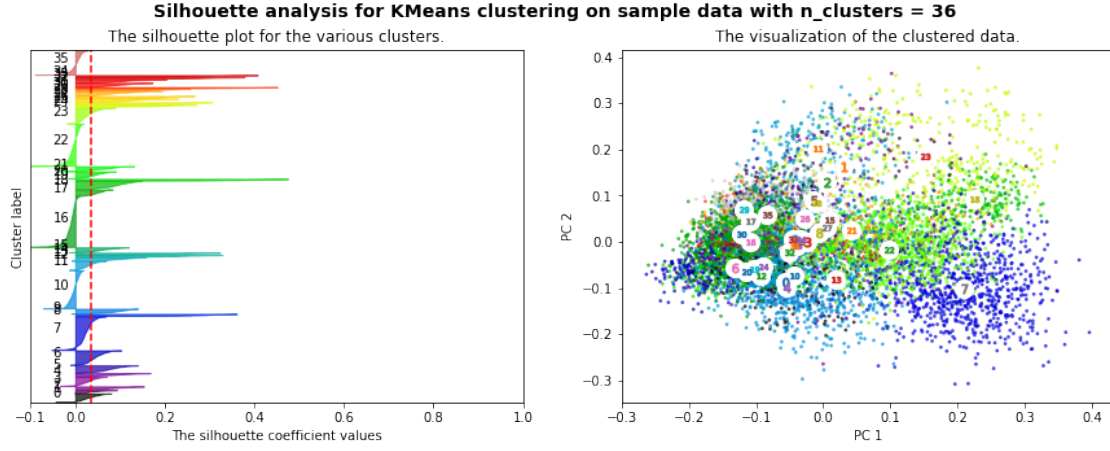
3.4. *K-means Clustering*

K-means clustering is an unsupervised learning method that partitions feature vectors into k clusters aiming at minimizing the sum of squared distance from each observation to the centroid with each cluster. We fit the model implemented by `sklearn.cluster.KMeans` using both the TF-IDF vectors and the Doc2Vec vectors as the features.

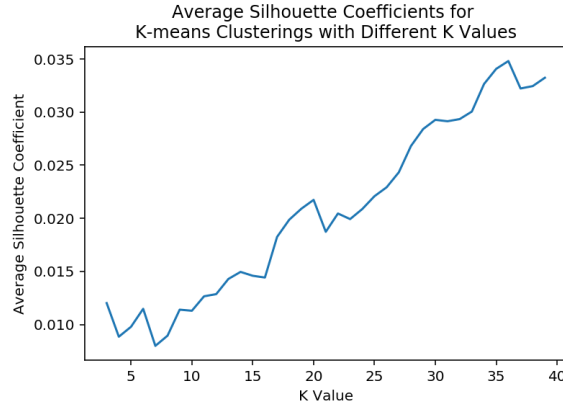
Fitting with the TF-IDF vectors, we cluster the single documents featured by distributions of normalized word frequency. We assume documents with similar TF-IDF vectors are likely to elaborate state positions about similar political agendas. On this basis, we can reveal the topics within each cluster by querying the most distinguishing words, in terms of TF-IDF scores, of the cluster centroids. Although this analysis reveals very limited information about the preference for political agenda of any specific state, it can help us with an overall impression about what the topics are discussed in the UN General debate in the past half-century.

Using the state vectors in the Doc2Vec semantic space, we are to identify state blocs in the semantic space, respectively. Since each vector represents the content of all the statements of one state as a whole, partitioning such vectors can reveal the state blocs in terms of position proximity. We also experiment by fitting the model using a subset of the corpus during the Cold War period, trying to identify state blocs in the Cold War years.

To tune the parameter k , the number of clusters, we use Silhouette coefficient as a metric. The Silhouette is a measure of the extent to which an observation is similar to its cluster compared to it is to other clusters. The higher the Silhouette is, the more coherent and less separative the observations within each cluster are. For our clustering model using TF-IDF vectors, the optimal k value is 36 (shown in Figure 3). For the model using Doc2Vec vectors, the optimal k value is 5 (shown in Figure 4).



(a) Silhouette Coefficient for $k = 5$



(b) Silhouette Coefficients for different k values

Figure 3: Parameter tuning for K-means Clustering using Doc2Vec vectors

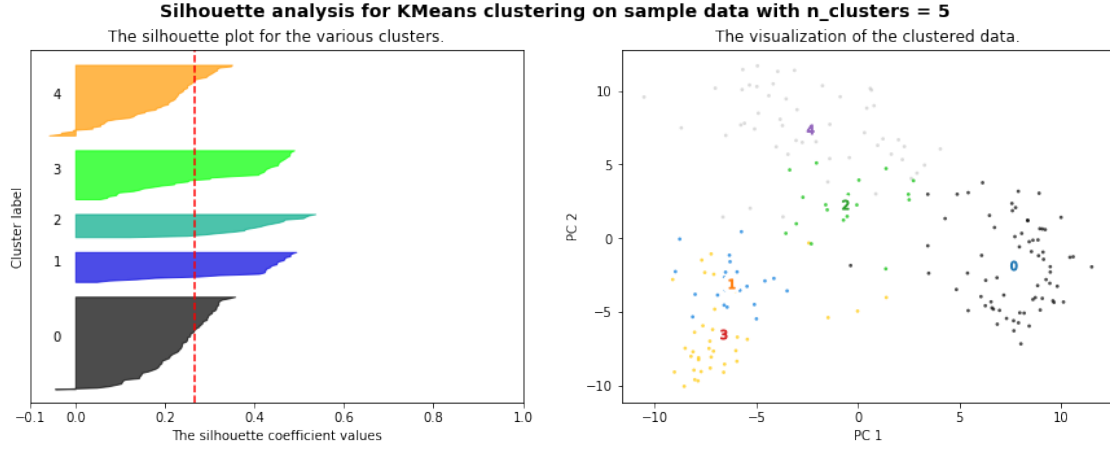
3.5. Topic Modeling

Topic Modeling is a method used for extracting abstract topics in a set of documents. In topic modeling, a document is represented by a distribution of topics, while a specific topic is represented by a probability distribution over a set of words. Words that are assigned with higher probability can better reflect the main ideas in this topic (Griffiths et al., 2007). In this paper, we use two different methods, The Latent Dirichlet Allocation model and Dynamic Topic Modeling¹, to address this issue from different perspectives.

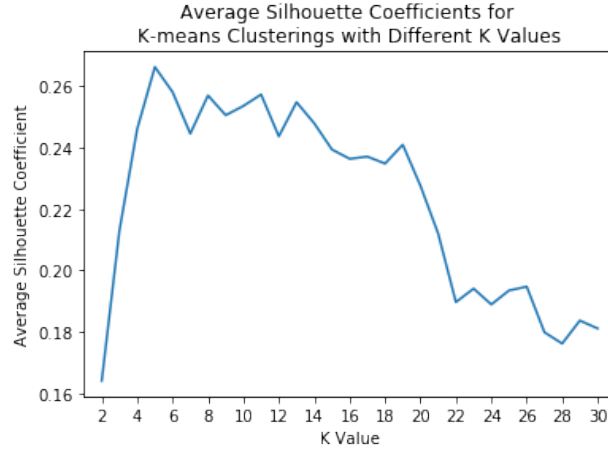
3.5.1. Latent Dirichlet Allocation (LDA) Model

We reveal state preferences for political agendas by revealing the topics in the corpus. By examining the distribution of probabilities of a certain speech on different topics, we can discover what issues the states that proposed this speech is putting spotlights on. This information is then

¹Dynamic Topic Modeling is requires substantial computing, we only provide the code in our notebook.



(a) Silhouette Coefficient for $k = 5$



(b) Silhouette Coefficients for different k values

Figure 4: Parameter tuning for K-means Clustering using Doc2Vec vectors

utilized to construct network analysis in the following section. Furthermore, generalized from a larger perspective, we may conclude the most heated topics on UNGA throughout the past half-century.

The Latent Dirichlet Allocation (LDA) model is one of the most widely used techniques for topic modeling. In this model, topics are presented as latent variables that cannot be observed. In the first two steps, θ_i , the topic distribution for the document i in the corpus and ϕ_k , the word distribution for the topic k , are characterized by two separate Dirichlet distributions at given Dirichlet prior parameters α and β . Then it supposes that $z_{i,j}$, the topic for the j -th word in document i can be expressed by a multinomial distribution of θ_i . Given the conditions above, a specific word at position j at a specific document i can, therefore, be represented by a multinomial distribution with the topics as an embedded layer (Blei et al., 2003).

The first step of LDA topic modeling in this paper is choosing the optimal parameters to obtain the most meaningful and interpretable topic results. Before conducting LDA analysis, we restrict

the vocabulary using TF-IDF measures to ignore words that appear too frequently in the corpus. In the TF-IDF measure, we also set up a tokenizer to stem words whose length is less than 4 letters. To acquire the optimal results of topics, we choose different levels of max_df ² parameter to compare the performance of the LDA model under different TF-IDF filtering standards.

The number of topics is another factor that we take into consideration when selecting the best LDA model. A satisfying topic number can capture the most information in the corpus and produces the most meaningful topic results. Given the two parameters of interest stated above, we use the UMass coherence score to evaluate the performance of the LDA model. UMass is an intrinsic measurement of topic coherence, which captures how well the topic model ‘explains’ our corpus of interest. The less the UMass coherence score is the better interpretability of the model. Figure 5 below illustrates the coherence score under different max_df level against different number of topics.

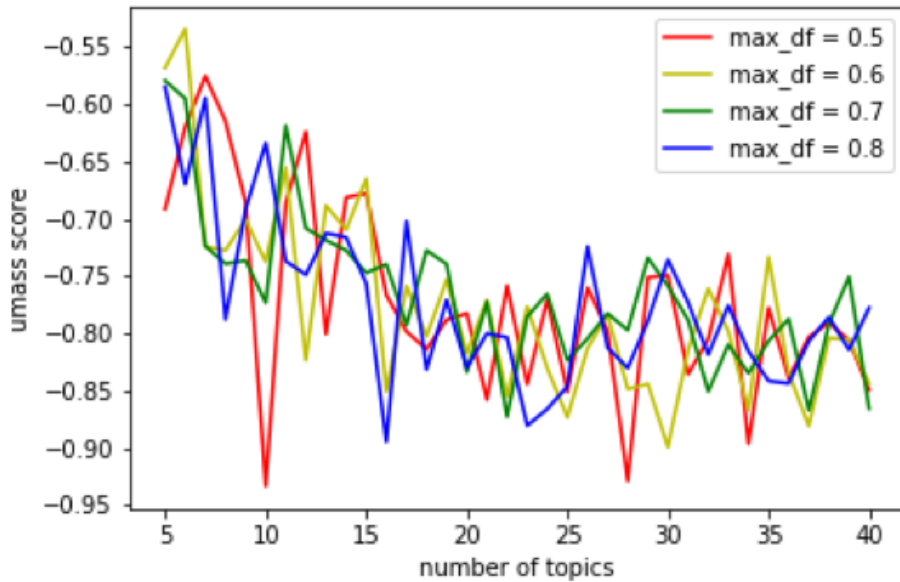


Figure 5: UMass coherence scores for different number of topics

As is shown in Figure 5, the LDA model has the best interpretability when setting $\text{max_df} = 0.5$ and choosing the number of topics equal to 10 or 28. After printing out the top feature words in each topic under these two topic numbers, we use $\text{num_topic} = 28$ for further analysis as it captures more information.

Using LDA topic modeling alone, we present the top 10 words with the highest probability in each topic. For some of these topics, we can interpret their practical contents. Besides, we use a Python package called `pyLDAvis` to visualize our topic model generated above. Other applications of the LDA model are stated in Section 3.6.

²A float between 0 and 1, which is used to indicate a proportion of documents, and words that have term frequency higher than that value is ignored.

3.6. Network Contraction: Mutual Information

Semantic networks are representations of semantic relationships between concepts of interest in the corpus. A semantic network has two key elements: nodes that represent concepts and edges, which can be directed or undirected, that represent the semantic relationship between linked concepts.

Based on the argument of Gurciullo and Mikhaylov (2017), we can regard topics extracted from the LDA model as having semantic values. Countries whose speeches cover similar topics show their overlying interest in similar political issues. In the last section, we use the LDA model to construct a vector that constitutes of topic probabilities of the 28 topics for each country each year. Therefore, in each of these 49 years, if we view each country (represented by a topic probability vector) as an individual semantic entity, we can build a semantic network based on the similarity of their topics of interest. The similarity is measured in normalized mutual information score, which will be introduced in detail in the following section.

In this paper, we construct 49 networks based on mutual information. We first investigate several statistics, which reveal the overall structure of the network such as density, average shortest length path, and diameter. Further on, by looking at the individual nodes of the graph, we can make inquiries about which countries are in the center of this network, which countries are on the border, and which countries are usually linked with one another. At the end of our network analysis, we go on to explore how a country's 'position' in the network evolves over years through different measurements of centrality.

3.6.1. Mutual Information

In the LDA model, for each country that has given a speech in each year, we obtain a vector of topic prevalence that describes the probability that the speech might load on each of the 28 topics. These can be intercepted as probability distributions, whose similarities between one another can be measured using the mutual information coefficient.

Mutual information captures the mutual dependence of two random variables. Since vectors here can be interpreted as discrete probability distributions, we might write the mutual information measurement as the following:

$$I_{X;Y} = \sum_{y \in Y} \sum_{x \in X} p_{X,Y}(x, y) \ln \frac{p_{X,Y}(x, y)}{p_X(x)p_Y(y)}$$

, where X and Y is the random variable of interest, $p_{X,Y}(x, y)$ is the joint probability density function of X and Y , and $p_X(x)$ and $p_Y(y)$ are the marginal probability density functions of X and Y , respectively.

The value of mutual information indicates how much the two random variables share information. In this case, a low mutual information score that close to zero denotes the two countries are focusing on different topics, while a relatively high mutual information score can be viewed as the two countries are paying attention to similar topics in that year. For comparison, we normalize the mutual information score computed above to a float between 0 and 1. The normalized mutual information score is what we used to construct edges in the semantic network.

3.6.2. Construct Networks using Mutual Information

After obtaining the normalized mutual information score between any two countries that have delivered a speech in a given year, we establish a semantic network whose nodes are countries. This network is completely linked at the first place, where the weight of the edge between two nodes is defined by the normalized mutual information score.

For the simplicity and interpretability of the network, we remove edges whose weights are below the median weight of the network. This step ensures that our networks only retain semantic relationships between countries above a certain threshold. Also, for years between 1970 and 1991, we assign different colors to nodes that represent NATO countries and WTO countries. An example of the network is shown in Figure 6.

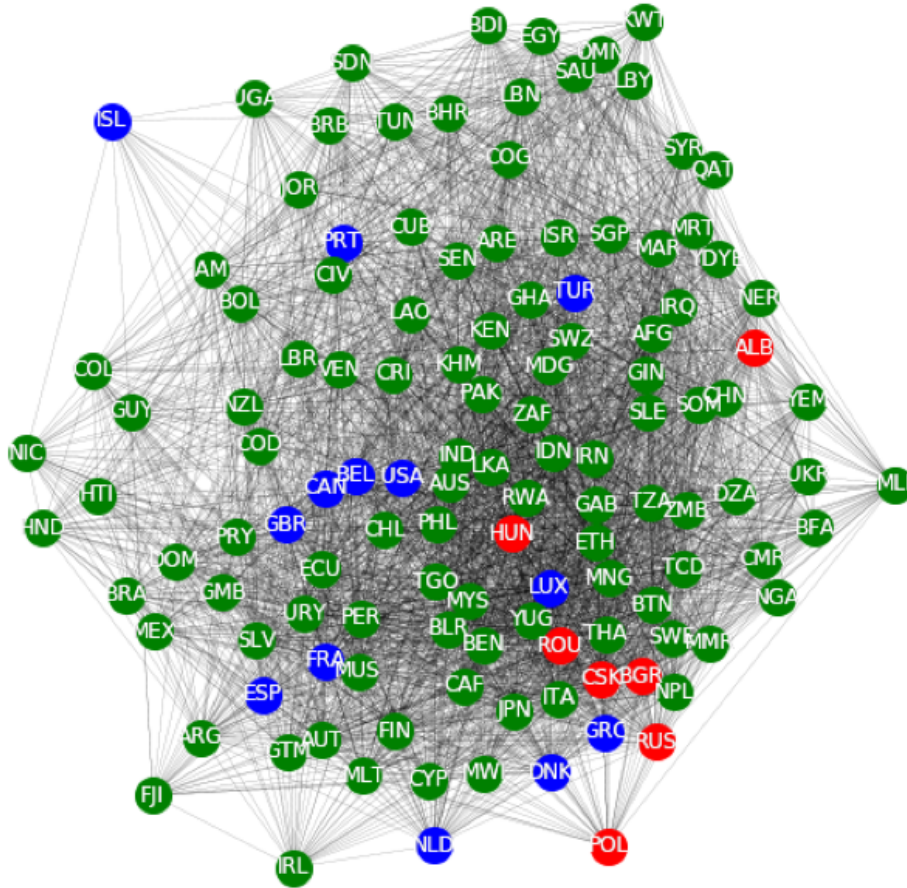


Figure 6: State Network of Mutual Information (1972; NATO states in blue and WTO states in red)

3.6.3. Network Statistics

The network statistics that we examine in this paper can be categorized into statistics that reveal general characteristics of the network and statistics that indicate features of individual nodes.

For the first category, we mainly examine density, average shortest path length, and diameter as indicated by Gurciullo and Mikhaylov’s paper in 2017. Density is a proportion of the actual number of edges over the total amount of possible edges. During the 49 years in our analysis, the average density is about 0.495, and its variance is about 0.004, which implies the network density is quite stable over the years.

As for average shortest path length and diameter, the networks of the years 1990 and 1993 do not have these features since these two networks are not connected. For other networks, the average shortest path length is about 1.508 and the average diameter is about 2.936.

For the second category, we mainly focus on degree centrality. Degree centrality indicates how ‘central’ a node is compared with other nodes, based on how many connections (edges) it has. Here in our analysis, degree centrality captures how similar a country’s speech is in topic comparing with the statements of other countries. The country with the highest degree centrality made a speech that covers the most heated topics discussed this year.

3.7. Topic summarization using BERT Text Generation

BERT (Bidirectional Encoder Representation from Transformers) is first introduced by Devlin et.al. in 2018[4]. Pre-training unlabeled texts from two directions (left and right) at all layers allows it to overcome the drawbacks of traditional unidirectional fine-tuning methods such as ELMo. This method has its unique advantages when conducting researches related to text generations and question answering.

Text generation has been widely used in indicative summary generations (Kan and Mckeown, 2002)[5], which is an unbiased way to derive the condensation of a corpus as well as extract its important features. In our analysis, we utilize this method to generate the summaries of certain topics of interest. Furthermore, through the automatic generation of texts, we try to examine the linguistic standard of a formal speech in the United Nations General Debate. To be specific, we endeavor to investigate what the topics are about in the language of the states addressing their opinions on UNGA.

4. Results

Employing the methods introduced in the previous section, we try to answer two research questions. First, we examine what topics states are concerned about and how the attention paid to specific issues evolve over the years. Second, we examine the patterns in which the states are related to each other in terms of topics of common concerns, political positions, and policy preferences, and the dynamics of the relationship.

4.1. State Preferences for Political Agenda

The most frequent words in this corpus hint us with an overall impression of the general topics being discussed on the UNGA in the past half-century. While, peace and development are the theme of the era, economic development, nuclear security, human rights, etc, are the common issues concerned by international society.

We try to gain further insights into agenda topics by performing K-means clustering on the TF-IDF matrix. We query the most distinguishing patterns of each cluster centroids, which is,

Table 4: Top 10 feature words in 28 topics

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 | Topic 8 | Topic 9 | Topic 10 |
|--|--|---|--|--|--|---|--|---|---|
| unite nation determination power self operation principle relation measure regime | unite nation sustainable challenge reform strengthen goal include terrorism address | unite nation union reform strengthen european secretary weapon programme challenge | unite nation afganistan principle affordable secretary lebanon terrorism challenge meet | nation unite sustainable challenge goal reform address climate strengthen secretary | unite nation live society fight women challenge cent child sustainable | unite nation small climate sustainable challenge tobago trinidad caribbean address | nation unite terrorist pakistan terrorism stand live attack want power | nation unite romania operation relation arab power principle regime mean | iraq unite nation iran arab iraqi palestinian principle kuwait relation |
| Topic 11 | Topic 12 | Topic 13 | Topic 14 | Topic 15 | Topic 16 | Topic 17 | Topic 18 | Topic 19 | Topic 20 |
| nation unite suriname poland principle panama base power weapon relation | nation unite sustainable include japan challenge live reform climate goal | unite nation mongolia relation principle increase initiative operation area reform | chad nation paraguay unite guinea live principle solidarity principe lead | unite syria israel nation palestinian terrorism syrian yemen arab terrorist | nation unite island climate pacific sustainable small challenge ocean address | unite nation sahel sudan mali tanzania lake niger hold refugee | unite nation ukraine weapon ukrainian russia central crimea principle strengthen | nation unite malawi weapon power canada delegation operation reform initiative | unite venezuela nation nicaragua bolivarian america latin american attempt sovereignty |
| Topic 21 | Topic 22 | Topic 23 | Topic 24 | Topic 25 | Topic 26 | Topic 27 | Topic 28 | | |
| unite nation viet china asean singapore power korea chinese korean | nation solomon unite nation tonga ireland weapon principle climate small | unite nation turkey leste timor greece determination european union portugal | unite nation mauritius philippines live increase ocean principle chagos lead | sudan unite nation lead drug libya terrorism democracy initiative arab | unite korea korean peninsula nation statement relation build joint bulgaria | morocoo nation unite albania mohammed kingdom king relation union european | malta nation unite mediterranean ethiopia power principle relation delegation operation | | |

Based on the feature words shown above, we can hand-classify these topics into 9 categories:

- **Climate Change (Topic 5, 7, 12, and 16):** These topics contain words such as “climate” and “sustainable”, which indicates a focus on climate change and sustainable development. Topic 5 and 12 discuss the general issues regarding climate change, while topic 7 and 16 put a special emphasis on marine countries.
- **Terrorism (Topic 4, 8, 10, 15 and 25):** The five topics all consist of keywords such as “terrorism” or “terrorist” while underscoring different perspectives individually. Topic 4 contains the word “afghanistan”, where the Taliban, one of the terrorist organization, is founded. Topic 8 is related to terrorism in Pakistan. Its northern part borders with Afghanistan, where terrorist activities are mostly conducted by religious extremists. Topic 10 focuses on terrorism in Iraq, which might have a relationship with Iraq War which started in 2003. Topic 15 is concerned with terrorism in Syria, which is still on the U.S. list of State Sponsors of Terrorism. Syria is accused of sponsoring weapons and providing political support for terrorists in Palestine and intervening in Iraq Conflicts, which corresponds to our results found in the LDA model. Topic 25 reveals terrorism issues in Sudan, where members of Hamas (Islamic

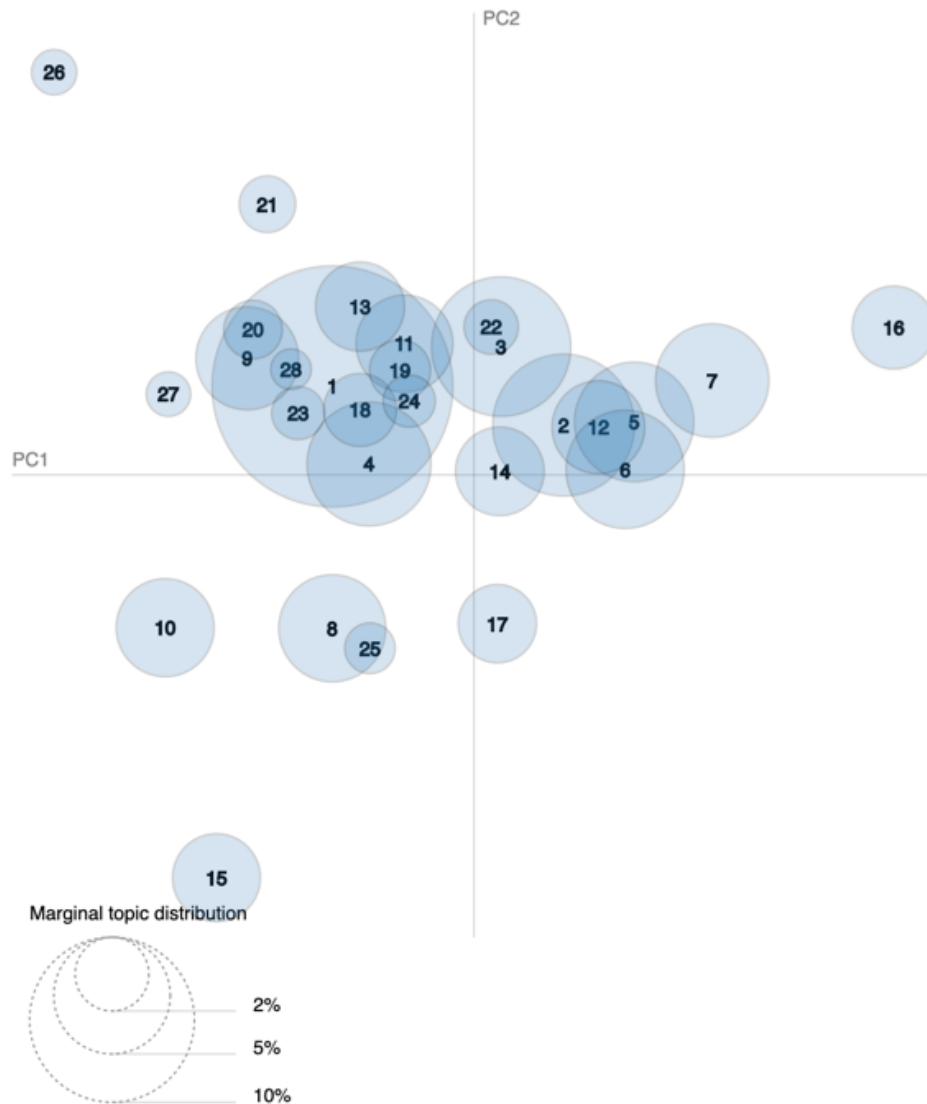


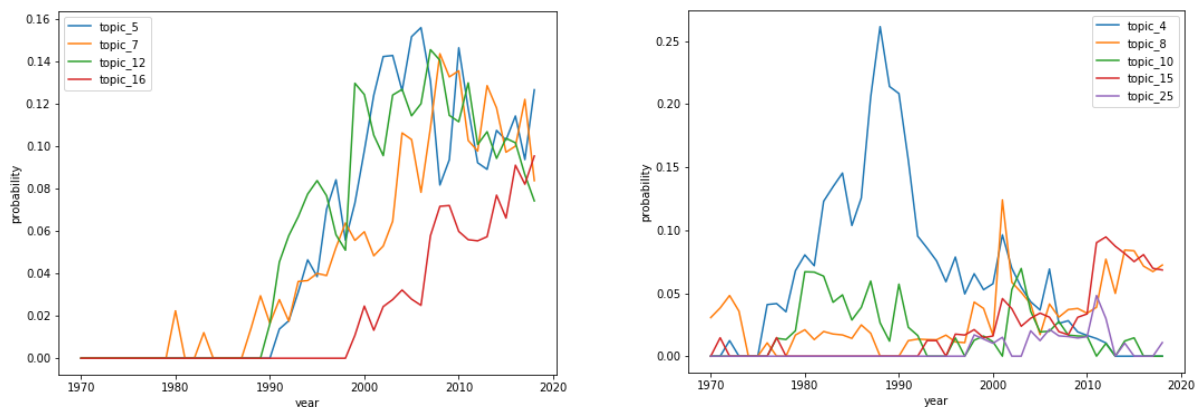
Figure 8: Intertopic Distance Map (after multidimensional scaling)

Resistance Movement) are allowed to live and fundraise.

- **Issues in Asian (Topic 21 and 26):** Topic 21 and 26 indicate issues in Asia. Topic 21 is related to general affairs in Asia since it incorporates words such as ‘viet’ (Vietnam), ‘china’ (chinese), ‘korea’ (korean), ‘singapore’, and ‘asean’ (Association of Southeastern Asian Nations). Topic 26 points to Korean issues in particular. This topic also includes other distinctive words like “peninsula”, “statement”, and “joint”, demonstrating the ongoing negotiation of reunification problem in this area.
- **Issues in Latin America (Topic 20):** Topic 20 deals with issues in Latin America, which

includes highly identifiable words such as ‘latin’, ‘america’ (american), along with country names such as ‘venezuela’ and ‘nicaragua’. There are still areas in Latin America whose sovereignty is still in dispute, which has been a subject discussed in the 1980s.

- **Issues in Europe (Topic 23):** Having words such as “european”, “union”, “greece”, and “turkey”, Topic 23 reveals a clear indication of its focus in European issues.
- **Issues in Africa (Topic 17):** Topic 17 addresses issues, especially refugee problems, in Africa. It also shows the most discussed countries in Africa, which are countries in Sahel, Sudan, Mali, and Tanzania.
- **Crimea Crisis (Topic 18):** Topic 18 mainly discusses war and conflicts in Crimea. After the dissolution of the Soviet Union, Crimea was formed as a constituent entity of Ukraine. After several conflicts, the sovereignty of Crimea was passed to the Russian Federation in 2014.
- **Human Rights (Topic 6):** Topic 6 differs from other topics in that it contains words like ‘women’, ‘child’, ‘society’, and ‘fight’, which might indicate its focus on human rights issues.
- **Others (Topic 1, 2, 3, 9, 11, 13, 14, 19, 22, 24, 27, and 28):** The 12 topics remaining cannot be easily categorized into any of the classifications above. They are also not identifiable enough to be regarded as an independent category. Some of these topics, such as topic 1, 2, 3, 9, and 13 include words that are mostly functional and do not have discernible political indications. Topic 11, 14, 19, 22, 24, 27 and 28 contain few incompatible feature words so that they cannot be assigned into any of the categories above.



(a) How topics on climate change have evolved over time (b) How topics on terrorism have evolved over time

Figure 9: Time trend of topics on certain issues

After deducing the meanings and categories of the 28 topics, we can examine how a topic’s probability has evolved by aggregating the texts of the same year. As is shown in Figure 9,

topics related to terrorism reveals different time patterns. Topic 4, which discusses terrorism in Afghanistan, attracts more global attention than other topics. This topic is at its peak around 1990. Issues about terrorism in Pakistan comes to a climax after 2000, and its probability has increased in current years after a partial decline. Topic 10 (terrorism in Iraq) has two peaks. One starts at 1980 and ends at 1990, and another starts after 2000, which correspond to the period of Iraq War. Topic 15 (terrorism in Syria) gains increased attention after 2010, while topic 25 (terrorism in Sudan) is the least discussed one among all issues about terrorism.

Table 5: Top 5 topics discussed over years (20 states)

| Country | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 |
|---------|-------|-------|-------|-------|-------|
| USA | 8 | 12 | 1 | 18 | 6 |
| CHN | 21 | 1 | 13 | 9 | 10 |
| JPN | 12 | 13 | 3 | 1 | 26 |
| IND | 1 | 4 | 8 | 5 | 3 |
| GBR | 1 | 8 | 12 | 4 | 2 |
| FRA | 1 | 3 | 6 | 13 | 14 |
| ITA | 22 | 3 | 1 | 2 | 18 |
| BRA | 11 | 1 | 6 | 3 | 5 |
| CAN | 19 | 2 | 12 | 1 | 3 |
| RUS | 18 | 13 | 1 | 3 | 26 |
| KOR | 5 | 26 | 2 | 12 | 13 |
| ESP | 1 | 14 | 3 | 11 | 22 |
| AUS | 3 | 1 | 2 | 16 | 4 |
| MEX | 11 | 6 | 1 | 3 | 2 |
| IDN | 4 | 5 | 1 | 3 | 24 |
| NLD | 2 | 1 | 3 | 19 | 18 |
| SAU | 10 | 4 | 15 | 1 | 8 |
| TUR | 23 | 4 | 1 | 15 | 2 |
| CHE | 2 | 6 | 12 | 3 | 28 |

Now we turn to state preferences revealed by LDA topic modeling. Table 5 illustrates the top 5 topics across all years for countries whose GDP ranks top 20 in 2019. The United States has a special interest in issues about terrorism, as they discuss terrorism in Pakistan more than any other topic, followed by topics about climate change, Crimea crisis, and human rights. China focuses on Asian issues (Topic 21). The result for all states is presented in Table A.13, A.14, A.15, A.16, A.17, and A.18.

We turn to the analysis based on the Doc2Vec semantic space. Recalling that we have labeled the documents with year and country code, we can examine the cosine similarity of the year vectors and other vectors representing specific issues. This allows us to reveal the dynamics of attention paid on the issues over the years. We choose the issues in Table 6 and select the corresponding keywords to construct the basis vectors. Note that our result is robust against slightly different keywords. We also measure the attention paid by the US and China on the same topics

by calculating the cosine similarity of the country vectors representing the US for each year and the basis vectors.

Table 6: Selected topics and corresponding key words

| Topic | Key words |
|----------------------|-----------------------|
| anti-terrorism | terrorism |
| nuclear weapons | nuclear |
| health | health |
| education | education |
| climate change | climate, environment |
| economic development | economic, development |



Figure 10: Dynamics of attentions on specific policy issues over years (1970-2018))

As is shown in Figure10, health, education, climate change, and economic development received substantial attention from the international society in recent decades. This can be largely explained by the efforts promoted by the United Nations on these issues. The graph shows a sharp increase in the international attention on terrorism in the year of 2001, suggesting the escalated concerns and tense discussions about this topic after the September 11 attack happened in New

York. As for the nuclear weapon issue, the international attention calmed down after 1991 when the Soviet Union collapsed and the Cold War terminates.

Among these topics, there is no significant emphasis put by the US or China. Even though the international society has shown constantly increasing concerns about health, education, climate change, and economic development, the attention paid by the US and China shift back and forth around zero cosine similarity on these topics. However, this may not necessarily suggest that the US and China do not prefer discussions or international efforts on such topics. On the contrary, this may be explained by the diversity of the statements made by the two countries each year. Due to time limitations on the General Debate section, states concerning a wide range of issues must have shortened the length of speech on specific topics. We can verify this by looking back at the original scripts of US statements. Nevertheless, it is still obvious that the dynamics of the US and China’s concerns on these topics are highly correlated to that of the international society, which indicates the significant role played by the two countries in the UNGA conference.

Finally, we present the result of topic summarization using BERT Text Generation. Table7 shows four examples of using BERT to generate texts based on General Debate Corpus. As we can see, formal speeches in the United Nations General Debate tend to focus on the global effect of a specific issue. It means a country is inclined to emphasize on how an issue will influence the world when talking about its topics of interest. Also, the countries also highlight a lot about the important position the United Nations is from a global perspective.

Table 7: Examples of topic summarization using BERT Text Generation

| Initial tokens | Generated Text |
|----------------------------|--|
| Climate change is... | Climate change is a major challenge for the world. The world is facing a new challenge, one that is not only global, but also global in scope. The United Nations is the only global organization that can effectively address the challenges of climate change. |
| Terrorism is... | Terrorism is a threat to international peace and security. It is a threat to the stability of the region and to the stability of the world. It is a threat to the stability of the world. |
| The Syrian conflict has... | The Syrian conflict has been a source of great concern to the international community. The Syrian people have suffered a terrible loss of life and have suffered a great loss of property. |

4.2. Policy Position Proximity and International Relations

Given the complexity of state policy objectives and international relations, it is beneficial to investigate state policy positions in high dimensions. We first assume that the policy positions are expressed in the General Debate statements and can be represented by the TF-IDF matrix. Then we turn to more sophisticated models to reveal similarities and discrepancies between state policy positions including the Doc2Vec semantic space and the LDA topic modeling.

To characterize high-dimensional state policy positions, we perform Principal Components Analysis on the TF-IDF matrix, which decomposes the TF-IDF vectors for each document to a

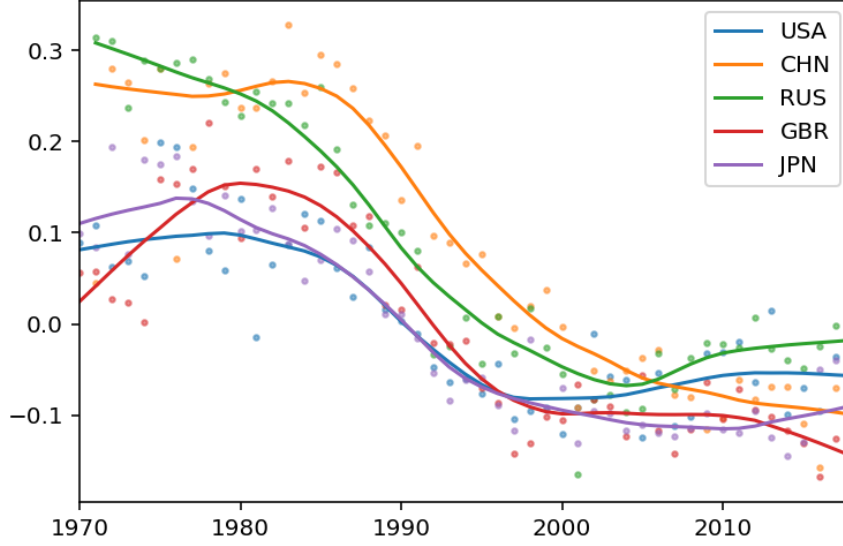


Figure 11: Multi-dimensional policy position revelation: principle component of the TF-IDF matrix (1970-2018))

scalar that accounts for the most variation in the data. In Figure 11, we visualize the results for the US, China, Russia (previously the Soviet Union), the UK, Japan, Korea, and the European Union to make a shallow observation of the international relations revealed by the TF-IDF matrix. It is obvious that, before 1990, China and Russia are similar to each other in terms of policy positions, and the US is closer to the UK and Japan. This is consistent with the historical context. While the Sino-soviet split from 1956 to 1966 is not captured by our data, the improvement of the bilateral diplomatic relationship after that period can be observed from the converging curves of China and the Soviet Union (labeled by RUS) in the figure. The consistency of our results with history increases our confidence in the validity of our analysis.

We investigate the policy position proximity during the Cold War period. This geopolitical tension mainly between the US and the Soviet Union spanned 45 years before its termination marked by the 1991 dissolution of the Soviet Union (Plan). We plot the PCA results for all the member states of the North Atlantic Treaty Organization (NATO) and the Warsaw Treaty Organization (WTO) in Figure 12. The NATO and WTO member states are labeled by the real lines and the dashed lines, respectively. It is a significant pattern that the NATO states and the WTO states show similar policy positions to each other within the organizations before the year of 1990. While the Soviet Union is located in the most extreme position in the figure among other WTO members, the US is laid in the middle of other allies in NATO. It is also obvious that the discrepancies decrease over years before the dissolution of the Soviet Union.

Furthermore, we can also conclude that the state policy positions have shown an overall trend of converging in recent decades, especially by the middle of the 2000s. However, discrepancies seem to have gradually escalated after 2008, the year of the international financial crisis.

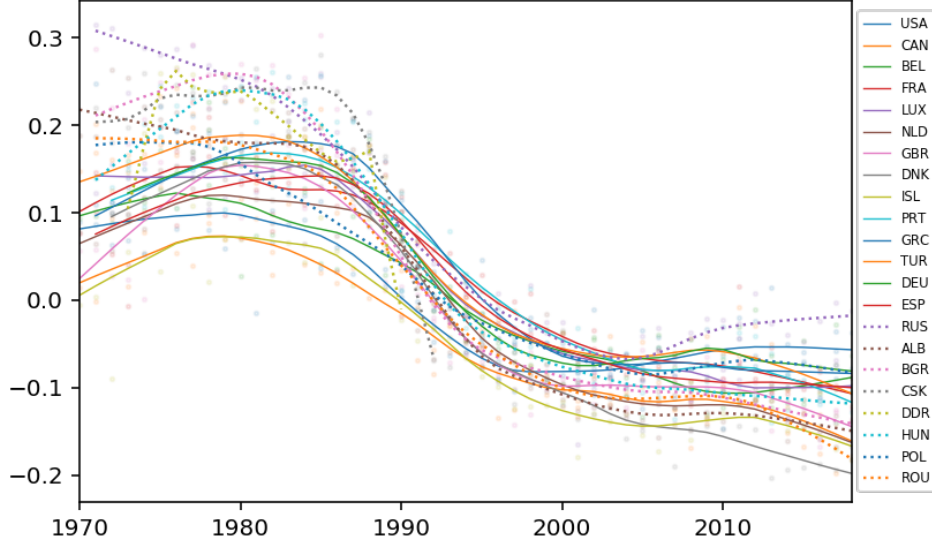


Figure 12: Multi-dimensional policy position revelation: NATO and WTO states (1970-2018)

We now turn to the analysis based on the Doc2Vec semantic space. We examine the cosine similarities between state vectors with the assumption that the policy positions are meaningfully underlying the vector representations of the document in the semantic space.

In Figure13, we visualize the pairwise cosine similarities of the 20 states that have the highest nominal GDP by 2020, according to the IMF. Red cells indicate significant dissimilarities between two states in terms of the meaning of their General Debate statement, and blue cells indicate similarities. Under our assumption, these are equivalent to dispute or disagreement and consensus or proximate policy positions, respectively, on policy issues. However, given that international relations can vary dramatically across time, merely the overall cosine similarities cannot capture the dynamics of policy position proximities.

To examine the dynamics of policy position proximities over years, we calculate the cosine similarities between the state vectors each year. As is shown in FigureA.16, we take the average of the results within every five years from 1970 to 2015 to cancel out noisy variations. The dynamics of recent years are shown individually in FigureA.17. We also visualize the standard deviations of the cosine similarities in FigureA.18.

Several observations can be made from the figures. First, the UK and the US are consistently similar to each other in terms of policy positions, which conforms to our findings in the previous analysis. Second, the cosine similarity between the US and China increase from 1971-1975 to 1976-1980. This can be explained by a series of historical events that mark the improvement of the US-China relations in the 1970s. Starting from 1971, the tension of the US-China relations had been alleviated by a series of civic exchange activities including the well-known Pingpong diplomacy. In February of 1972, President Nixon's visit to Beijing marks the resumption of the

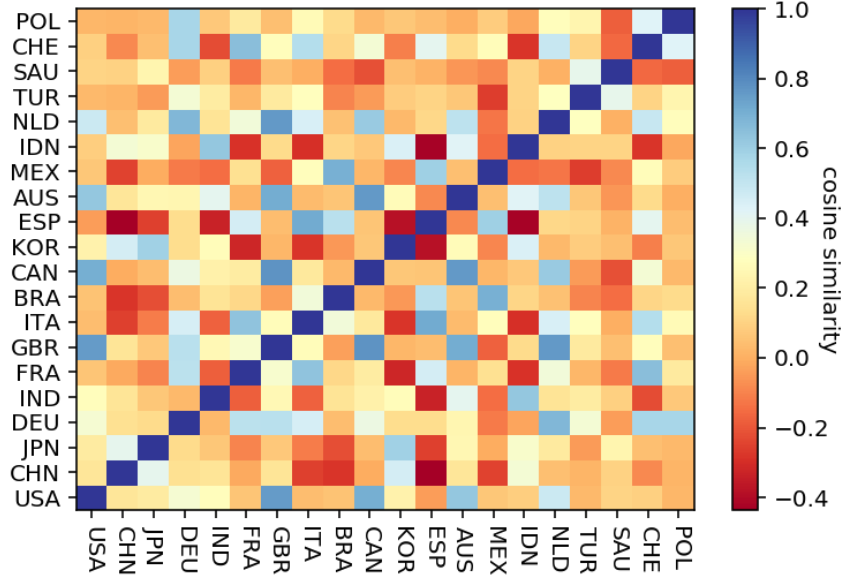


Figure 13: Pairwise cosine similarities of 20 states (1970-2018)

harmonious relations between the two countries, and the US officially changed its diplomatic recognition of China from Taipei to Beijing on January 1, 1979. Besides, there are also significant results that can hardly be explained by history. For example, while the figure suggests significant dissimilarities between Spain and French, and French and India, we cannot match any important historical events to the analytical results.

Again, we apply this analysis to examine the Cold War period. We train the Doc2Vec model again using the 1970-1991 subset of the UNGDC corpus (see FigureA.20). In FigureA.19, all the states in our corpus are projected onto a basis vector constructed by the difference of the mean of the NATO state vectors and the WTO state vectors as well as the tokens "capitalism" and "socialism". NATO and WTO member states are indicated by bigger labels. Blue labels indicate proximity to the NATO camp, and red labels indicate proximity to the WTO camp. Only the related positions on the north-to-south direction are meaningful, and horizontal variations are randomly assigned to make a clear visualization. This result is a simple prediction of whether a state is closer to the NATO camp or to the WTO camp before the dissolution of the Soviet Union.

We finally perform K-means clustering on the Doc2Vec state vectors. The result is shown in 14. After tuning the parameter, the optimal number of clusters is 5, which means the states can be optimally partitioned into five state blocs. As is shown in TableA.19, the top states listed in cluster 0, Australia (AUS), Canada (CAN), the UK (GBR), Ghana (GHA), the Gambia (GMB) are members of the Commonwealth of Nations. The top states in cluster 2 are mostly South American countries, and cluster 3 mainly involves middle eastern countries. The result is correlated to the status quo of the geopolitical structure of the world. It identifies latent state blocs in terms of policy position proximities.

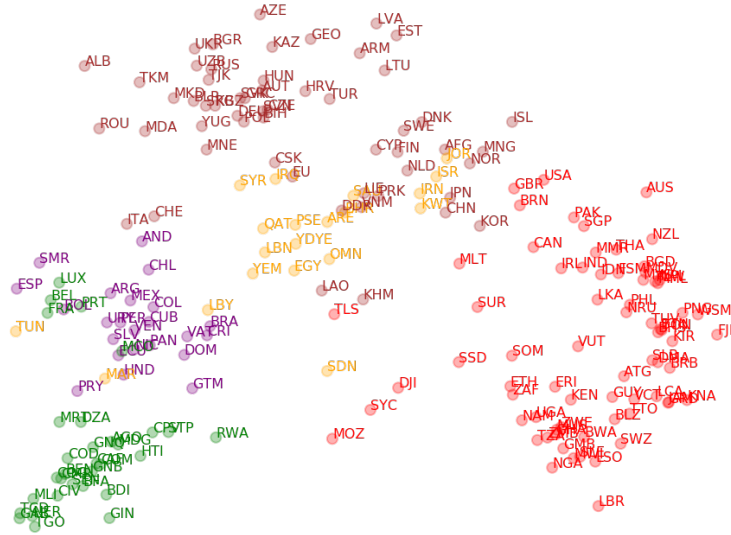


Figure 14: K-means clustering of Doc2Vec state vectors

All 49 networks constructed by normalized mutual information score is presented in the appendix due to the limitation of space in the main body. By examining networks from 1970 to 1991, we can discover that NATO countries and WTO countries are generally close to countries of the same kind, even though countries of the same organization may not necessarily gather in a cluster. Moreover, there is not much distance between NATO countries and WTO countries. There are two possible explanations for this phenomenon. The first one is that there are indeed few divergences in opinions about topics between these countries. As we can see from our analysis on topic modeling in the former section, the most discussed topics across all years are those about climate change, terrorism, and other functional topics with no intelligible political implications. Possibilities are strong that these countries agree to each other on these overall topics in spite of their difference in ideology. Another plausible explanation of the proximity is that mutual information only captures the similarity between topics included in two countries' speeches, not the opinions underlying these speeches. Therefore, chances still exist that countries with different ideologies take different stands on political issues, even though they are focused on the same topic.

Another insight worth noting is that networks of the year 1990 and 1993 have subgraphs that are not connected. In the network of the year 1990, the country East Germany is an isolated node. In 1993, the isolated node is Cambodia. It implies that these two countries talk about topics differed from all other countries in that year.

Now we turn to network statistics mentioned in Section 3.6.3. Figure 15 below illustrates

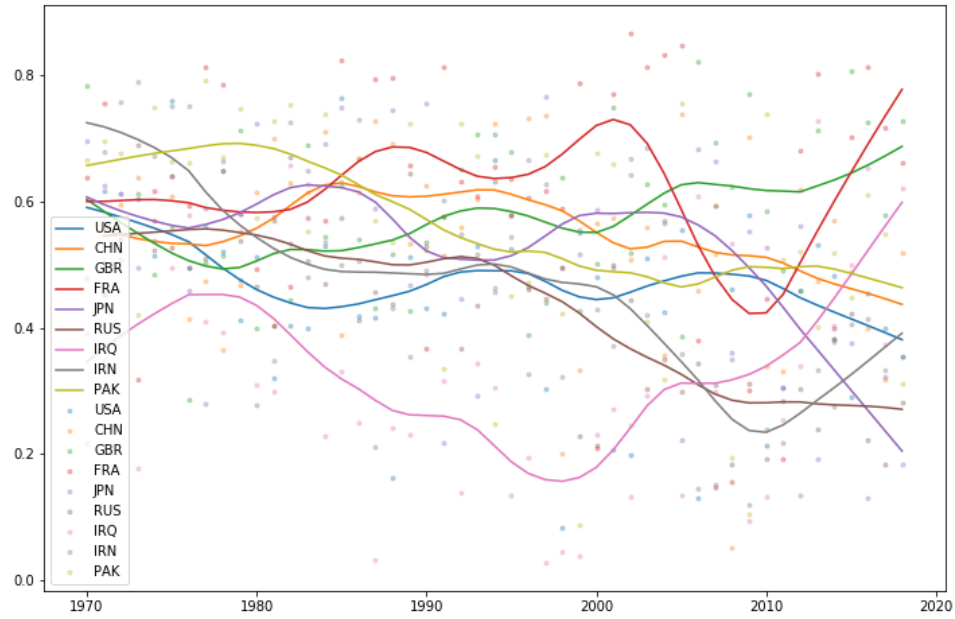


Figure 15: Degree centrality from 1970 to 2018

how several countries' degree centrality has changed over time. As introduced in the former section, degree centrality captures how similar a country's speech is compared with those of other countries in topic distribution. It is shown in the figure that France generally has the highest degree centrality score across the 50 years, and Iraq has the lowest. For most of the countries in the figure above, their degree centralities have decreased since 2000. Generally speaking, as demonstrated in the graph, developed countries (here we have USA, GBR, GBR, JPN) are more central than developing countries and third World countries (here we have CHN, RUS, IRA, IRN, PAK) in networks.

5. Reflections and Extensions

K-means clustering using TF-IDF. No time slice, features are associated with single document, which is the distinguishing features of the specific state in the specific year.

References

- Baturo, A., Dasandi, N., Mikhaylov, S.J., 2017. Understanding state preferences with text as data: Introducing the un general debate corpus. *Research & Politics* 4, 2053168017712821.
- Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, 993–1022.
- Griffiths, T.L., Steyvers, M., Tenenbaum, J.B., 2007. Topics in semantic representation. *Psychological review* 114, 211.
- Gurciullo, S., Mikhaylov, S., 2017. Topology analysis of international networks based on debates in the united nations. *arXiv preprint arXiv:1707.09491*.

- Jankin Mikhaylov, S., Baturo, A., Dasandi, N., 2017. United Nations General Debate Corpus. URL: <https://doi.org/10.7910/DVN/0TJX8Y>, doi:10.7910/DVN/0TJX8Y.
- Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-sne. *Journal of machine learning research* 9, 2579–2605.
- Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Plan, M., . Marshall plan. Public law 80, 472.

Appendix A. Tables and Graphs

Table A.8: Changes of participating states over years

| year | plus | minus |
|------|---|---|
| 1971 | IRL LUX POL CHL MLI CHN YDYE CAF BFA MUS PAN M... | HND CRI GMB |
| 1972 | MWI HND SWZ CRI DNK ARE BTN GMB BHR OMN BRB PRT | PAN NOR TTO |
| 1973 | PAN DEU BHS LSO DDR NOR | MWI DOM SWZ GMB CIV UGA TGO MLT CMR TCD BRB |
| 1974 | DOM SWZ GMB BGD GRD CIV GNQ UGA BWA TGO MLT CM... | PAN FJI ZAF BHS MDG |
| 1975 | MOZ PAN MWI FJI MDG | LBN LUX SWZ GMB HTI CIV EGY MLT |
| 1976 | LBN CPV SWZ MDV PNG COM GNB STP EGY MLT SUR | MWI GNQ THA |
| 1977 | MWI LUX AGO WSM GMB TTO HTI CIV SYC THA | GUY DOM GAB MLT |
| 1978 | DOM GNQ GAB VNM MLT GUY | MWI SWZ PRY GIN SAU |
| 1979 | DJI MWI DMA PRY VAT GIN SAU | GMB DNK SWE AGO |
| 1980 | SWZ LCA VCT AGO DNK SWE GMB BHS ZWE | KWT WSM DMA VAT |
| 1981 | KWT WSM DMA | NOR MWI SWZ MMR GMB TTO LSO |
| 1982 | NOR ATG MMR GMB TTO BLZ LSO | LBN DMA WSM CIV SYC |
| 1983 | LBN VUT SLB WSM SYC | PAN LCA GMB |
| 1984 | PAN SWZ LCA DMA GMB ZAF | VCT WSM GRD SYC ZMB |
| 1985 | MWI WSM BRN KNA | LBN CPV ATG SEN LCA DMA MDV BTN GMB GNB FJI GN... |
| 1986 | LBN CPV ATG SEN VCT MDV BTN GMB GNB GRD FJI GN... | MDG UGA BEN KNA |
| 1987 | BEN LCA KNA GAB MDG | STP ATG |
| 1988 | STP ATG UGA | GUY |
| 1989 | GUY CIV SYC DMA | LBN SLE WSM BTN CAF |
| 1990 | LBN SLE WSM BTN LIE CAF NAM | YDYE STP KHM CIV |
| 1991 | EST KOR PRK LTU LVA FSM CIV STP MHL KHM | COD DDR NER SOM |
| 1992 | TJK BIH KGZ MDA HRV SVN NER GEO SMR ARM UZB AZ... | LCA WSM TTO CAF KHM ZMB GUY YUG THA |
| 1993 | LCA WSM TKM MKD CZE TTO CAF SVK KHM ERI GUY MC... | VUT GEO STP CSK MDG |
| 1994 | GEO ZAF ZMB MDG AND | WSM TKM |
| 1995 | WSM TKM | CPV DMA COM CIV UZB SYC ZWE SAU |
| 1996 | CPV VUT DMA COM PLW CIV UZB SYC ZWE SAU | CMR |
| 1997 | | WSM PLW FSM KHM SAU |
| 1998 | WSM FSM PSE STP CMR SAU | MHL |
| 1999 | PLW KHM MHL | CAF SYC STP |
| 2000 | CAF NRU SOM | MOZ WSM BTN CMR UZB ZMB |
| 2001 | MOZ WSM TON BTN STP ZMB CMR UZB TUV SYC YUG | |
| 2002 | CHE | LBY SYC |
| 2003 | KIR TLS SYC | DJI TKM |
| 2004 | DJI VAT TKM LBY | SOM |
| 2005 | SOM | DJI HND MLI LBR BWA SYC SAU OMN |
| 2006 | SRB HND MNE MLI LBR BWA SYC SAU OMN | YUG |
| 2007 | | SAU MLI |
| 2008 | MLI | |
| 2009 | DJI | |
| 2010 | | DJI MDG UZB TKM |
| 2011 | DJI SSD TKM UZB EU MDG | SYC |
| 2012 | SAU SYC | GNB |
| 2013 | GNB | DJI KEN SAU |
| 2014 | KEN | |
| 2015 | DJI SAU | SGP CMR UZB |
| 2016 | SGP CMR UZB | DJI BRN |
| 2017 | DJI BRN | |
| 2018 | | |

Table A.9: Distinguishing features of K-means cluster centroids

| Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 | Cluster 8 |
|------------|-------------|-------------|-------------|--------------|--------------|-------------|-----------|-----------|
| iraq | terrorism | soviet | morocco | bahamas | africa | guatemala | sudan | ecuador |
| kuwait | reform | socialist | algeria | drug | south | belize | malawi | peru |
| lebanon | cooperation | nuclear | mauritania | haiti | african | guatemalan | somalia | latin |
| iraqi | nuclear | republic | arab | caribbean | namibia | american | ethiopia | peruvian |
| lebanese | afghanistan | disarmament | maghreb | small | independence | central | kenya | american |
| arab | weapon | relation | african | traffic | regime | guatemalans | eritrea | america |
| iran | millennium | weapon | sahara | island | struggle | america | somali | democracy |
| aggression | terrorist | imperialist | africa | illicit | apartheid | caribbean | africa | drug |
| israel | poverty | detente | kingdom | commonwealth | delegation | drug | african | andean |
| islamic | sustainable | union | palestinian | south | republic | democracy | igad | power |

Table A.10: Distinguishing features of K-means cluster centroids (continued)

| Cluster 9 | Cluster 10 | Cluster 11 | Cluster 12 | Cluster 13 | Cluster 14 | Cluster 15 | Cluster 16 | Cluster 17 |
|-------------|--------------|------------|---------------|------------|-------------|------------|------------|-------------|
| japan | south | african | korea | chad | azerbaijan | latin | guinea | sustainable |
| nuclear | delegation | africa | lao | chadian | armenia | chile | equatorial | climate |
| japanese | africa | congo | korean | african | karabakh | argentina | papua | agendum |
| assistance | operation | republic | republic | libya | armenian | venezuela | bissau | goal |
| operation | nuclear | burundi | peninsula | africa | azerbaijani | brazil | pacific | tobago |
| weapon | power | reform | democratic | libyan | nagorny | mexico | african | woman |
| reform | disarmament | democratic | nuclear | republic | nagorno | american | republic | trinidad |
| disarmament | relation | liberia | north | delegation | minsk | america | africa | 2015 |
| korea | namibia | continent | cooperation | idriiss | osce | uruguay | south | poverty |
| intend | independence | delegation | reunification | darfur | armenians | power | delegation | mdgs |

Table A.11: Distinguishing features of K-means cluster centroids (continued)

| Cluster 18 | Cluster 19 | Cluster 20 | Cluster 21 | Cluster 22 | Cluster 23 | Cluster 24 | Cluster 25 | Cluster 26 |
|-------------|------------|------------|-------------|-------------|------------|-------------|-------------|-------------|
| european | island | ukraine | operation | israel | ireland | arab | panama | latvia |
| kosovo | marshall | georgia | europe | arab | irish | yemen | canal | baltic |
| union | barbuda | ukrainian | nuclear | palestinian | northern | terrorism | panamanian | reform |
| bosnia | antigua | georgian | european | israeli | nuclear | emirate | latin | latvian |
| herzegovina | small | abkhazia | disarmament | palestine | european | bahrain | american | european |
| europe | pacific | russian | weapon | jordan | operation | egypt | panamanians | sustainable |
| cooperation | caribbean | russia | negotiation | territory | unionist | palestinian | america | afghanistan |
| moldova | climate | european | south | aggression | british | syrian | treaty | union |
| croatia | nuclear | europe | soviet | occupy | weapon | stability | republic | europe |
| albania | test | chernobyl | arm | zionist | violence | israeli | oceanic | syria |

Table A.12: Distinguishing features of K-means cluster centroids (continued)

| Cluster 27 | Cluster 28 | Cluster 29 | Cluster 30 | Cluster 31 | Cluster 32 | Cluster 33 | Cluster 34 | Cluster 35 |
|------------|-------------|---------------|-------------|------------|------------|---------------|--------------|------------|
| turkey | pacific | spain | pakistan | paraguay | caribbean | tunisia | myanmar | honduras |
| cyprus | island | spanish | india | bolivia | saint | arab | drug | nicaragua |
| turkish | fiji | gibraltar | sri | bolivian | barbados | tunisian | rakhine | costa |
| greece | zealand | operation | lanka | latin | grenada | maghreb | narcotic | haiti |
| greek | solomon | european | kashmir | america | small | african | opium | salvador |
| cypriots | small | mediterranean | nuclear | american | island | solidarity | poppy | rica |
| cypriot | climate | europe | afghanistan | drug | dominica | palestinian | nuclear | dominican |
| european | samoa | possible | terrorism | democracy | lucia | mediterranean | reform | central |
| settlement | tuvalu | crisis | indian | coca | kitts | ben | constitution | el |
| union | sustainable | terrorism | south | paraguayan | haiti | stability | democratic | american |

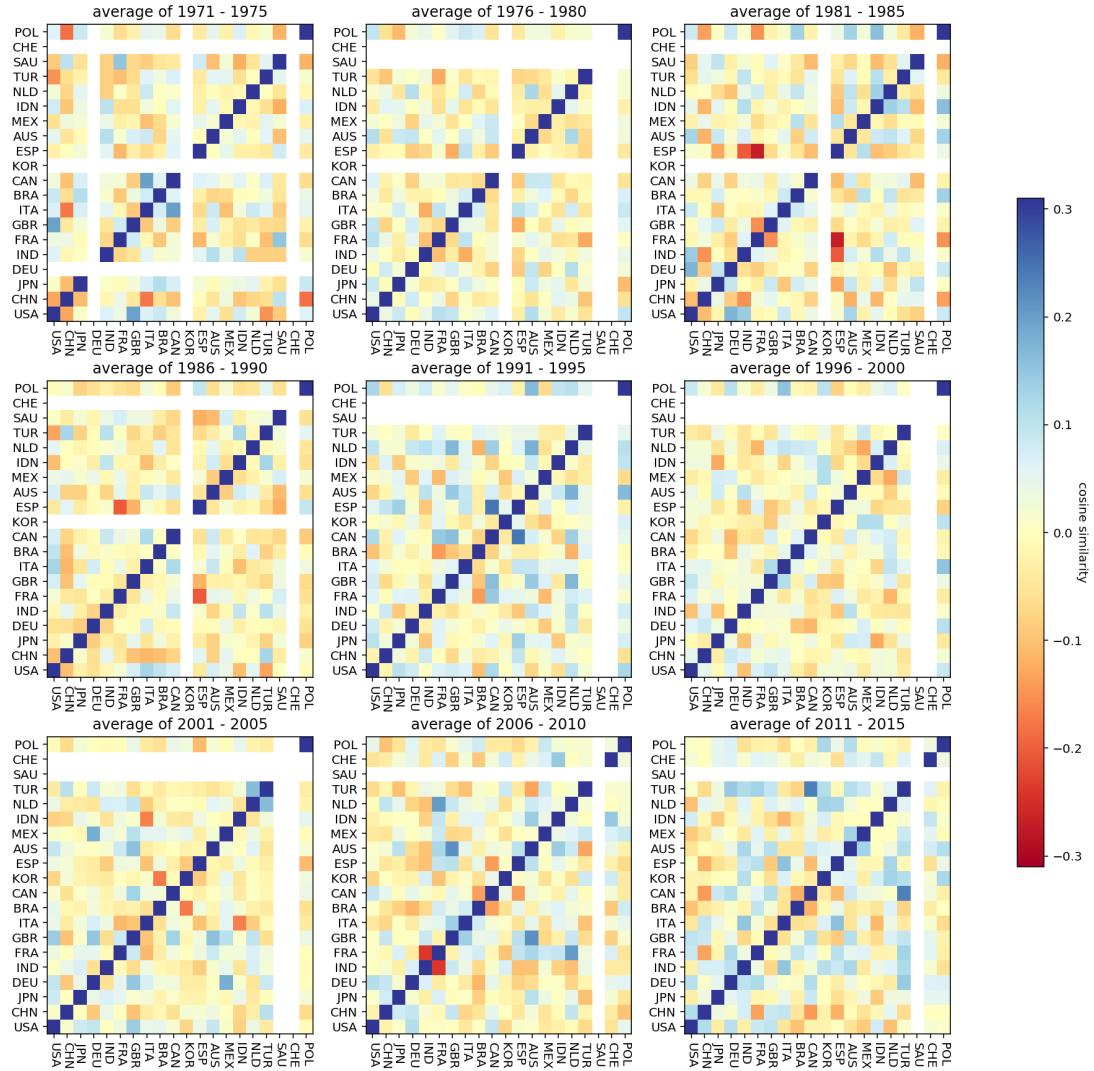


Figure A.16: Average pairwise cosine similarities of 20 states every five years (1970-2015)

Table A.13: Top 5 topics discussed over years

| Country | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 |
|---------|-------|-------|-------|-------|-------|
| AFG | 4 | 8 | 1 | 18 | 3 |
| AGO | 1 | 17 | 23 | 3 | 9 |
| ALB | 27 | 1 | 9 | 18 | 8 |
| AND | 6 | 12 | 2 | 8 | 1 |
| ARE | 10 | 4 | 15 | 1 | 5 |
| ARG | 11 | 1 | 20 | 14 | 6 |
| ARM | 2 | 8 | 1 | 23 | 3 |
| ATG | 7 | 1 | 6 | 8 | 12 |
| AUS | 3 | 1 | 2 | 16 | 4 |
| AUT | 2 | 13 | 3 | 18 | 22 |
| AZE | 2 | 13 | 10 | 18 | 15 |
| BDI | 1 | 17 | 6 | 3 | 14 |
| BEL | 3 | 1 | 18 | 2 | 6 |
| BEN | 1 | 6 | 3 | 14 | 5 |
| BFA | 1 | 6 | 5 | 17 | 3 |
| BGD | 4 | 3 | 5 | 24 | 7 |
| BGR | 3 | 26 | 2 | 18 | 13 |
| BHR | 10 | 15 | 4 | 1 | 12 |
| BHS | 7 | 1 | 4 | 2 | 3 |
| BIH | 2 | 8 | 3 | 1 | 18 |
| BLR | 1 | 13 | 3 | 9 | 18 |
| BLZ | 7 | 16 | 1 | 6 | 4 |
| BOL | 6 | 14 | 25 | 11 | 1 |
| BRA | 11 | 1 | 6 | 3 | 5 |
| BRB | 7 | 4 | 1 | 12 | 16 |
| BRN | 12 | 5 | 1 | 4 | 13 |
| BTN | 19 | 5 | 7 | 1 | 4 |
| BWA | 1 | 5 | 4 | 7 | 12 |
| CAF | 1 | 14 | 3 | 17 | 18 |
| CAN | 19 | 2 | 12 | 1 | 3 |
| CHE | 2 | 6 | 12 | 3 | 28 |
| CHL | 11 | 1 | 6 | 3 | 14 |
| CHN | 21 | 1 | 13 | 9 | 10 |
| CIV | 1 | 3 | 14 | 6 | 17 |
| CMR | 1 | 14 | 3 | 19 | 4 |
| COD | 1 | 17 | 6 | 14 | 10 |

Table A.14: Top 5 topics discussed over years (continued)

| Country | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 |
|---------|-------|-------|-------|-------|-------|
| COG | 1 | 17 | 6 | 14 | 3 |
| COL | 6 | 11 | 1 | 25 | 14 |
| COM | 14 | 1 | 10 | 4 | 24 |
| CPV | 1 | 14 | 6 | 5 | 3 |
| CRI | 11 | 6 | 1 | 12 | 14 |
| CUB | 1 | 20 | 6 | 8 | 21 |
| CYP | 1 | 23 | 4 | 2 | 13 |
| CZE | 2 | 18 | 3 | 12 | 13 |
| DEU | 3 | 18 | 13 | 1 | 2 |
| DJI | 1 | 4 | 17 | 15 | 7 |
| DMA | 7 | 16 | 12 | 4 | 1 |
| DNK | 2 | 3 | 18 | 1 | 4 |
| DOM | 14 | 1 | 6 | 11 | 7 |
| DZA | 1 | 27 | 10 | 5 | 4 |
| ECU | 11 | 6 | 1 | 14 | 12 |
| EGY | 15 | 1 | 10 | 4 | 25 |
| ERI | 1 | 8 | 17 | 28 | 7 |
| ESP | 1 | 14 | 3 | 11 | 22 |
| EST | 2 | 18 | 12 | 3 | 8 |
| ETH | 1 | 28 | 17 | 5 | 4 |
| EU | 2 | 8 | 6 | 15 | 18 |
| FIN | 2 | 3 | 1 | 19 | 18 |
| FJI | 12 | 16 | 1 | 22 | 7 |
| FRA | 1 | 3 | 6 | 13 | 14 |
| FSM | 16 | 12 | 22 | 7 | 3 |
| GAB | 1 | 14 | 3 | 17 | 5 |
| GBR | 1 | 8 | 12 | 4 | 2 |
| GEO | 2 | 8 | 18 | 6 | 13 |
| GHA | 1 | 4 | 7 | 5 | 12 |
| GIN | 1 | 14 | 17 | 3 | 5 |
| GMB | 4 | 7 | 1 | 5 | 12 |
| GNB | 14 | 1 | 3 | 5 | 17 |
| GNQ | 14 | 1 | 3 | 6 | 28 |
| GRC | 23 | 1 | 3 | 2 | 4 |
| GRD | 7 | 1 | 6 | 4 | 8 |
| GTM | 11 | 14 | 6 | 3 | 1 |

Table A.15: Top 5 topics discussed over years (continued)

| Country | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 |
|---------|-------|-------|-------|-------|-------|
| GUY | 1 | 7 | 4 | 3 | 16 |
| HND | 18 | 6 | 14 | 11 | 1 |
| HRV | 2 | 1 | 18 | 3 | 13 |
| HTI | 1 | 6 | 7 | 14 | 4 |
| HUN | 3 | 1 | 18 | 13 | 2 |
| IDN | 4 | 5 | 1 | 3 | 24 |
| IND | 1 | 4 | 8 | 5 | 3 |
| IRL | 22 | 1 | 2 | 4 | 8 |
| IRN | 10 | 8 | 1 | 4 | 18 |
| IRQ | 10 | 1 | 8 | 15 | 28 |
| ISL | 24 | 2 | 1 | 12 | 16 |
| ISR | 8 | 15 | 1 | 10 | 18 |
| ITA | 22 | 3 | 1 | 2 | 18 |
| JAM | 7 | 1 | 3 | 4 | 16 |
| JOR | 15 | 1 | 4 | 10 | 8 |
| JPN | 12 | 13 | 3 | 1 | 26 |
| KAZ | 2 | 13 | 3 | 18 | 12 |
| KEN | 1 | 17 | 4 | 7 | 3 |
| KGZ | 2 | 3 | 13 | 5 | 6 |
| KHM | 13 | 21 | 1 | 8 | 24 |
| KIR | 16 | 12 | 7 | 5 | 6 |
| KNA | 7 | 12 | 4 | 6 | 5 |
| KOR | 5 | 26 | 2 | 12 | 13 |
| KWT | 10 | 4 | 1 | 15 | 8 |
| LAO | 21 | 1 | 9 | 5 | 3 |
| LBN | 4 | 15 | 8 | 1 | 10 |
| LBR | 1 | 5 | 3 | 12 | 7 |
| LBY | 1 | 25 | 10 | 15 | 8 |
| LCA | 7 | 1 | 12 | 8 | 4 |
| LIE | 2 | 1 | 12 | 18 | 3 |
| LKA | 4 | 1 | 8 | 3 | 7 |
| LSO | 1 | 4 | 5 | 7 | 19 |
| LTU | 2 | 18 | 3 | 12 | 8 |
| LUX | 3 | 1 | 2 | 18 | 15 |
| LVA | 2 | 18 | 3 | 12 | 5 |
| MAR | 27 | 1 | 10 | 4 | 15 |

Table A.16: Top 5 topics discussed over years (continued)

| Country | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 |
|---------|-------|-------|-------|-------|-------|
| MCO | 2 | 6 | 3 | 7 | 12 |
| MDA | 2 | 3 | 18 | 13 | 5 |
| MDG | 1 | 3 | 6 | 24 | 4 |
| MDV | 7 | 4 | 1 | 24 | 5 |
| MEX | 11 | 6 | 1 | 3 | 2 |
| MHL | 16 | 12 | 22 | 7 | 2 |
| MKD | 2 | 23 | 3 | 1 | 6 |
| MLI | 1 | 17 | 3 | 6 | 4 |
| MLT | 28 | 1 | 2 | 3 | 12 |
| MMR | 3 | 4 | 1 | 5 | 13 |
| MNE | 2 | 5 | 3 | 28 | 28 |
| MNG | 13 | 3 | 5 | 2 | 9 |
| MOZ | 1 | 5 | 9 | 17 | 3 |
| MRT | 1 | 25 | 10 | 17 | 4 |
| MUS | 24 | 1 | 5 | 7 | 4 |
| MWI | 19 | 5 | 7 | 12 | 3 |
| MYS | 4 | 8 | 1 | 12 | 5 |
| NAM | 5 | 1 | 7 | 4 | 17 |
| NER | 1 | 17 | 3 | 14 | 6 |
| NGA | 1 | 5 | 4 | 3 | 7 |
| NIC | 20 | 6 | 1 | 11 | 14 |
| NLD | 2 | 1 | 3 | 19 | 18 |
| NOR | 2 | 3 | 1 | 4 | 19 |
| NPL | 5 | 4 | 1 | 19 | 13 |
| NRU | 16 | 12 | 7 | 2 | 28 |
| NZL | 16 | 1 | 12 | 13 | 3 |
| OMN | 10 | 4 | 15 | 1 | 24 |
| PAK | 8 | 4 | 1 | 13 | 3 |
| PAN | 11 | 6 | 1 | 20 | 2 |
| PER | 6 | 11 | 3 | 1 | 14 |
| PHL | 24 | 1 | 12 | 5 | 6 |
| PLW | 12 | 16 | 7 | 28 | 28 |
| PNG | 16 | 12 | 1 | 3 | 14 |
| POL | 11 | 3 | 13 | 1 | 2 |
| PRK | 26 | 13 | 1 | 12 | 21 |
| PRT | 1 | 23 | 3 | 2 | 14 |

Table A.17: Top 5 topics discussed over years (continued)

| Country | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 |
|---------|-------|-------|-------|-------|-------|
| PRY | 14 | 1 | 6 | 11 | 19 |
| PSE | 15 | 8 | 12 | 28 | 28 |
| QAT | 10 | 15 | 1 | 4 | 12 |
| ROU | 9 | 3 | 2 | 13 | 1 |
| RUS | 18 | 13 | 1 | 3 | 26 |
| RWA | 1 | 17 | 6 | 14 | 3 |
| SAU | 10 | 4 | 15 | 1 | 8 |
| SDN | 25 | 1 | 4 | 15 | 10 |
| SEN | 1 | 6 | 14 | 3 | 17 |
| SGP | 1 | 12 | 21 | 8 | 7 |
| SLB | 22 | 16 | 12 | 7 | 5 |
| SLE | 1 | 5 | 16 | 4 | 7 |
| SLV | 6 | 1 | 11 | 14 | 3 |
| SMR | 2 | 6 | 3 | 5 | 12 |
| SOM | 1 | 17 | 8 | 15 | 12 |
| SRB | 2 | 6 | 27 | 8 | 12 |
| SSD | 25 | 17 | 12 | 5 | 2 |
| STP | 1 | 14 | 6 | 5 | 12 |
| SUR | 11 | 7 | 3 | 1 | 4 |
| SVK | 2 | 3 | 18 | 13 | 12 |
| SVN | 2 | 3 | 1 | 12 | 18 |
| SWE | 2 | 3 | 1 | 18 | 19 |
| SWZ | 5 | 12 | 1 | 4 | 19 |
| SYC | 1 | 16 | 24 | 6 | 7 |
| SYR | 15 | 1 | 10 | 8 | 4 |
| TCD | 14 | 1 | 17 | 25 | 10 |
| TGO | 1 | 3 | 14 | 17 | 6 |
| THA | 7 | 4 | 21 | 1 | 5 |
| TJK | 3 | 2 | 13 | 4 | 5 |
| TKM | 13 | 2 | 3 | 28 | 28 |
| TLS | 23 | 6 | 12 | 5 | 8 |
| TON | 16 | 22 | 12 | 5 | 7 |
| TTO | 7 | 1 | 4 | 3 | 5 |
| TUN | 1 | 10 | 15 | 3 | 5 |
| TUR | 23 | 4 | 1 | 15 | 2 |
| TUV | 16 | 12 | 5 | 7 | 28 |

Table A.18: Top 5 topics discussed over years (continued)

| Country | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 |
|---------|-------|-------|-------|-------|-------|
| TZA | 1 | 17 | 5 | 4 | 7 |
| UGA | 1 | 12 | 17 | 4 | 5 |
| UKR | 18 | 13 | 3 | 9 | 1 |
| URY | 1 | 11 | 6 | 3 | 14 |
| USA | 8 | 12 | 1 | 18 | 6 |
| UZB | 2 | 3 | 13 | 4 | 8 |
| VAT | 6 | 2 | 12 | 1 | 8 |
| VCT | 7 | 1 | 8 | 4 | 6 |
| VEN | 6 | 1 | 20 | 11 | 8 |
| VNM | 21 | 1 | 5 | 13 | 3 |
| VUT | 16 | 1 | 12 | 8 | 7 |
| WSM | 16 | 1 | 12 | 4 | 7 |
| YEM | 15 | 10 | 1 | 4 | 9 |
| ZAF | 1 | 5 | 12 | 6 | 3 |
| ZMB | 1 | 5 | 4 | 9 | 7 |
| ZWE | 1 | 5 | 4 | 7 | 8 |

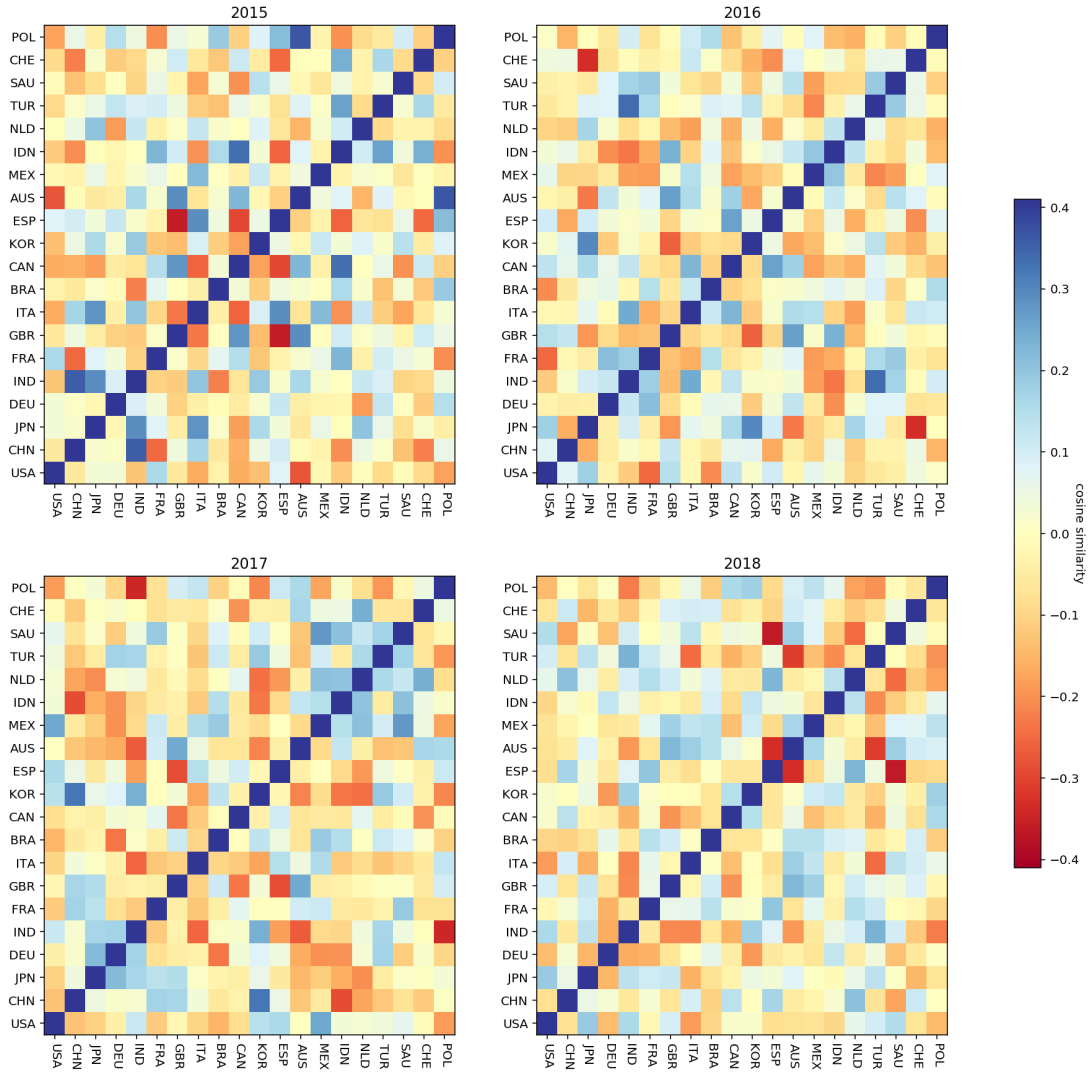


Figure A.17: Pairwise cosine similarities of 20 states in recent years (2015-2018)

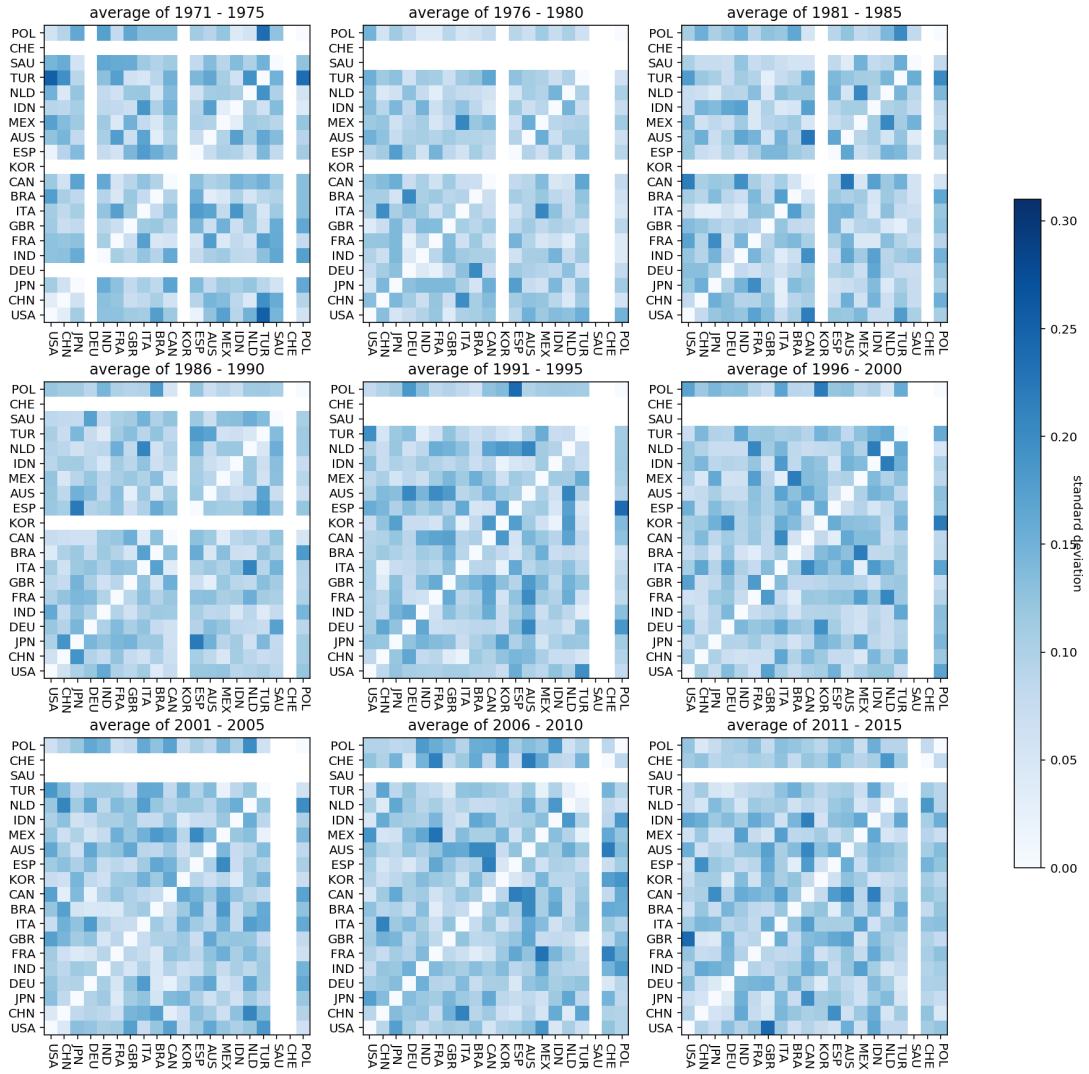


Figure A.18: Standard deviation of pairwise cosine similarities of 20 states every five years (1970-2015)

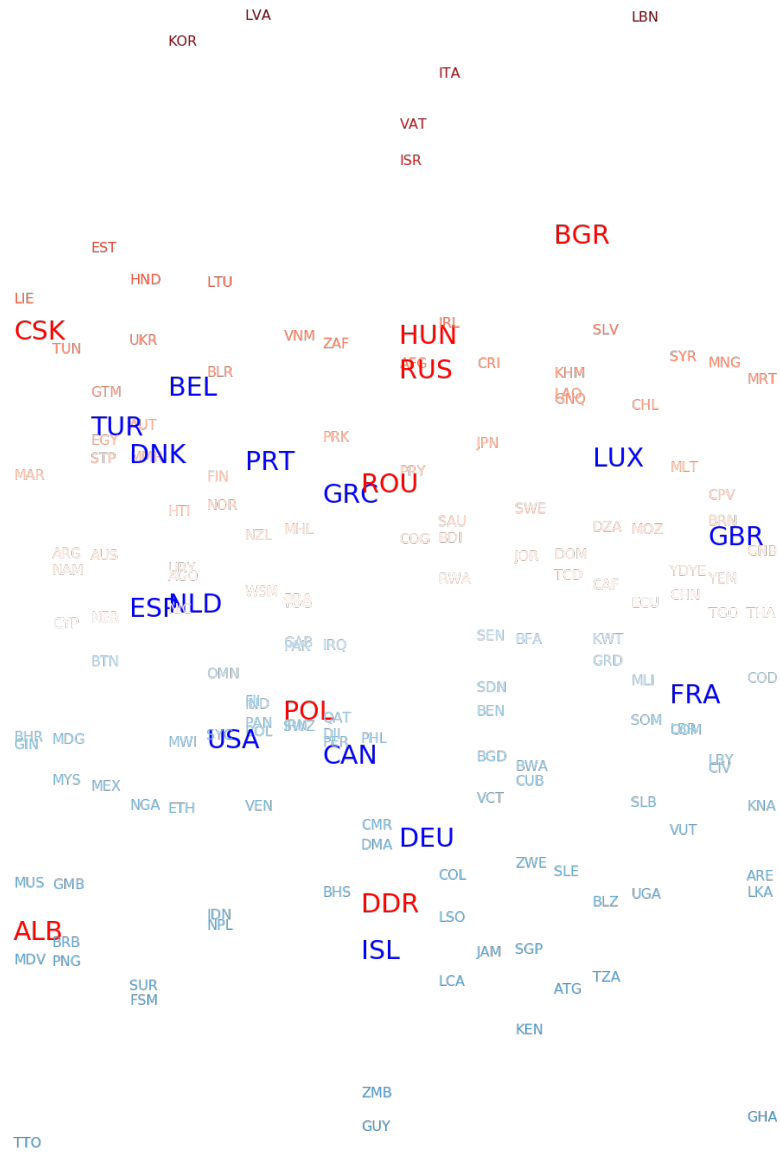


Figure A.19: Projection of states vectors onto the Cold War dimension (1970-1991)

Cold War: Semantic Space of Document Vectors
 (All nations and statements of 1970-1991,
 TSNE dimension reduction,
 Doc2Vec trained on subset of 1970-1991)

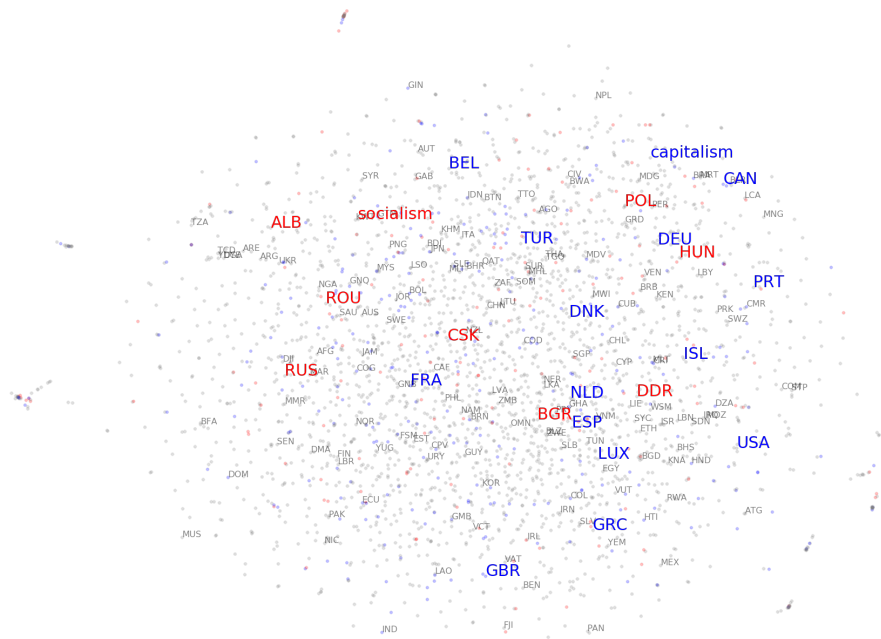


Figure A.20: Semantic space of document vectors (1970-1991)

Table A.19: K-means clustering of Doc2Vec state vectors

| cluster 0 | cluster 1 | cluster 2 | cluster 3 | cluster 4 |
|-----------|-----------|-----------|-----------|-----------|
| AUS | ALB | ARG | IRN | BEL |
| CAN | AUT | BOL | IRQ | CMR |
| GBR | BLR | BRA | ISR | COG |
| GHA | ISL | COL | KWT | DZA |
| GMB | ITA | CRI | LBN | FRA |
| IDN | JPN | CUB | LBY | GIN |
| IND | KHM | DOM | MAR | HTI |
| KEN | NLD | ECU | SDN | MDG |
| LBR | NOR | GTM | SYR | RWA |
| LKA | TUR | HND | TUN | TGO |
| MMR | UKR | MEX | EGY | BDI |
| NZL | YUG | PER | JOR | BEN |
| PAK | AFG | PRY | QAT | BFA |
| PHL | BGR | SLV | SAU | CAF |
| SGP | CHN | URY | YDYE | CIV |
| SLE | CSK | VEN | YEM | COD |
| SOM | CYP | CHL | ARE | GAB |
| THA | FIN | ESP | BHR | LUX |
| TTO | GRC | NIC | OMN | MLI |
| USA | HUN | PAN | PSE | MRT |
| ZAF | LAO | VAT | | NER |
| ZMB | MNG | SMR | | SEN |
| ETH | POL | AND | | TCD |
| FJI | ROU | | | PRT |
| GUY | RUS | | | GNQ |
| IRL | SWE | | | COM |
| JAM | DNK | | | CPV |
| MLT | DDR | | | GNB |
| MUS | DEU | | | STP |
| MYS | VNM | | | AGO |
| NGA | LIE | | | MCO |
| NPL | EST | | | |
| TZA | KOR | | | |
| UGA | LTU | | | |
| BRB | LVA | | | |
| BTN | PRK | | | |
| MWI | ARM | | | |
| ... | ... | | | |