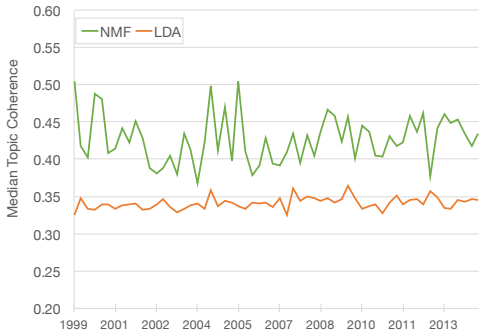
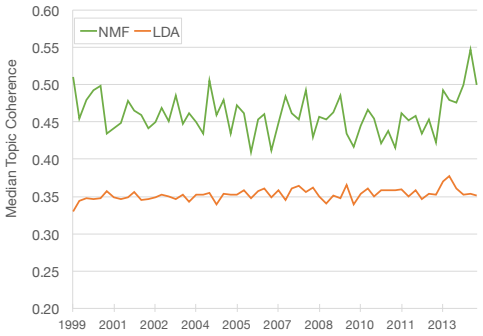
Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach (Supplementary Material)

Appendix A: Baseline Comparison

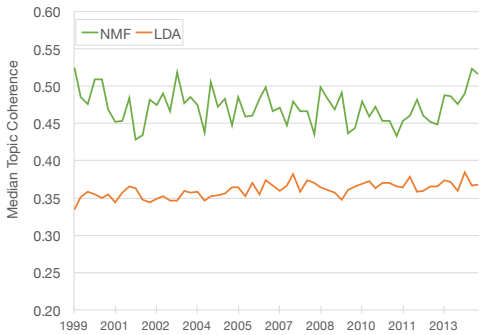
Here we compare NMF and LDA when applied for static topic modeling. Figures 1 and 2 show a full comparison of the coherence of the models generated by NMF and LDA on 60 time window datasets, for numbers of topics $k \in [10, 50]$.



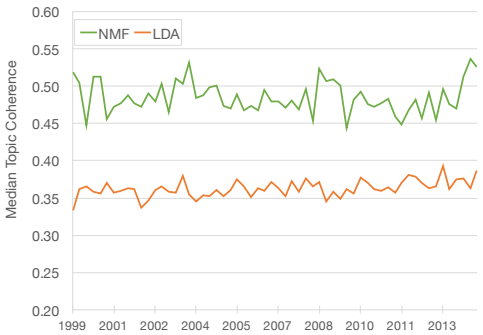
(a) C_v at $k = 10$ topics



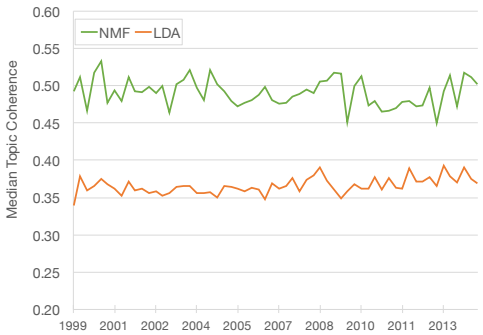
(b) C_v at $k = 20$ topics



(c) C_v at $k = 30$ topics

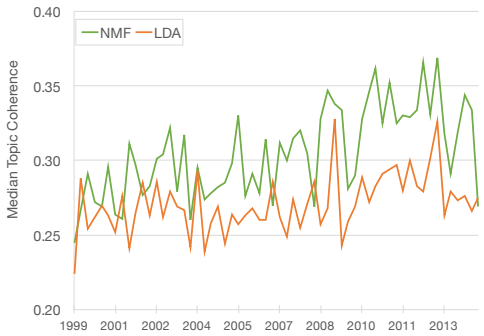


(d) C_v at $k = 40$ topics

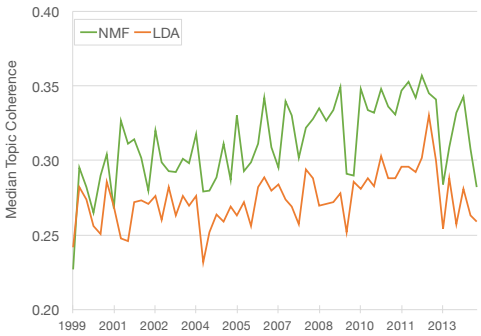


(e) C_v at $k = 50$ topics

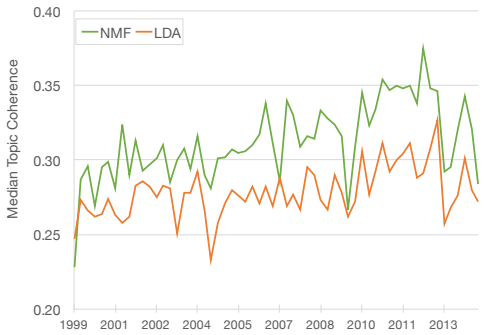
Figure 1: Median C_v topic coherence scores for all 60 time window datasets, for models produced by NMF and LDA with $k \in [10, 50]$ topics.



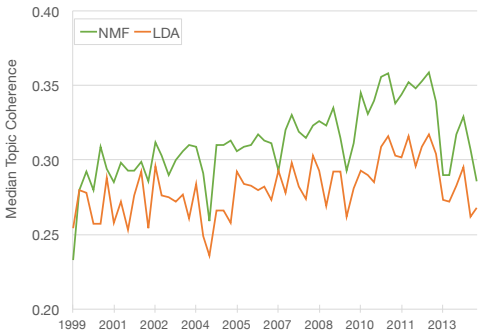
(a) TC-W2V at $k = 10$ topics



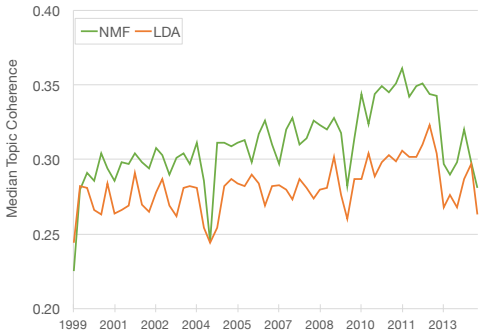
(b) TC-W2V at $k = 20$ topics



(c) TC-W2V at $k = 30$ topics



(d) TC-W2V at $k = 40$ topics



(e) TC-W2V at $k = 50$ topics

Figure 2: Median TC-W2V topic coherence scores for all 60 time window datasets, for models produced by NMF and LDA with $k \in [10, 50]$ topics.

Appendix B: Topic Model Validation

Intra-Topic Validity

To examine the intra-topic semantic validity of the dynamic topics produced by our approach, we examined the distribution of TC-W2V coherence values for all dynamic topics, when evaluated in the *word2vec* space built from the complete speech corpus. These coherence values correspond to the mean of the pairwise cosine similarities between the top-10 terms for each topic in the *word2vec* space. As evidenced by the coherence values reported in Table 1, the most coherent topics often correspond to core EU competencies. Unsurprisingly, broad administrative topics prove to be least coherent (*e.g.* 'Commission questions', 'Council Presidency', 'Plenary administration'). Overall the mean topic coherence score of 0.36 is considerably higher than the lower bound for TC-W2V (*i.e.* minimum value $= -1$), suggesting a high level of semantic validity.

Inter-Topic Validity

To assess the inter-topic semantic validity of the results, we examine the extent to which any meaningful higher-level grouping exists among the 57 dynamic topics. To do this we apply average linkage agglomerative clustering to the topics. Using the approach described in Greene et al. (2008), we re-cluster the row vectors from the second-layer NMF factor $\mathbf{H}$ using normalized Pearson correlation as a similarity metric. Here the vectors correspond the weights of each dynamic topic with respect to the 2,710 terms noted above. The dendrogram for the hierarchical clustering is shown in Fig. 3. Following the interpretation provided in Quinn et al. (2010), the lower the height at which any two topics are connected in the dendrogram, the more similar their term usage patterns in EP sessions.

We observe a number of higher-level groupings of interest, which are highlighted in Fig. 3. These includes groups specifically related to transport, ('Transport', 'Air transport', 'Maritime issues', 'Road safety') energy ('Climate change', 'Energy', 'Nuclear proliferation'), animal health ('Drugs', 'Foot & mouth', 'Animal health and welfare'), interactions with other institutions ('Council Presidency program', 'Council Presidency', 'Commission questions'), Education and research ('Education and Culture', 'Research'), trade ('Trade with China', 'Trade partnerships', 'WTO & aid'), and EU enlargement ('Enlargement', 'Turkish accession', 'The Balkans'). These hierarchical relationships between topics provide semantic validity for the model presented, where topics we would expect to be related are found to be correlated in the NMF factor $\mathbf{H}$ (*i.e.* they share similar

Topic	Short Label	Top 10 Terms	Coh.	Freq.
13	Transport	transport, railway, rail, passenger, road, network, freight, system, train, infrastructure	0.54	19
42	The Balkans	kosovo, serbia, balkan, resolution, bosnia, albania, iceland, herzegovina, macedonia, process	0.50	12
33	Air transport	air, passenger, transport, aviation, airport, traffic, airline, flight, sky, single	0.48	10
29	Adjusting to globalisation	fund, globalisation, egf, worker, adjustment, mobilisation, european, redundant, application, eur	0.47	15
6	Energy	energy, gas, renewable, efficiency, supply, source, electricity, market, target, project	0.47	36
39	Education & culture	programme, education, culture, language, cultural, youth, sport, learning, young, training	0.43	21
8	Fisheries	fishery, fishing, fish, stock, fisherman, fleet, sea, common, policy, measure	0.43	34
2	Human rights	rights, human, fundamental, freedom, democracy, law, charter, resolution, union, violation	0.43	52
45	Maritime issues	port, sea, maritime, safety, ship, accident, oil, vessel, transport, inspection	0.43	10
21	Healthcare	health, patient, environment, safety, public, care, healthcare, action, disease, mental	0.42	18
26	Child protection	child, internet, pornography, sexual, school, exploitation, young, victim, education, crime	0.42	14
56	Road safety	road, safety, vehicle, transport, system, driver, accident, motor, noise, ecall	0.41	12
16	Research	research, programme, innovation, framework, funding, industry, technology, development, cell, institute	0.41	15
15	Turkish accession	turkey, turkish, accession, progress, cyprus, negotiation, union, membership, croatia, macedonia	0.41	20
35	Tax	tax, vat, taxation, rate, system, fraud, states, evasion, car, transaction	0.41	11
32	Trade - WTO & aid	trade, wto, world, development, developing, international, negotiation, aid, free, relation	0.39	19
47	Product labelling & regulation	product, medicinal, medicine, tobacco, labelling, safety, consumer, regulation, organic, advertising	0.39	11
11	Trade - Trade partnerships	agreement, partnership, morocco, trade, negotiation, data, cooperation, association, korea, fishery	0.39	18
49	Regional funds	policy, region, cohesion, development, regional, strategy, structural, fund, economic, area	0.39	22
17	CFSP	security, policy, defence, common, foreign, military, nato, immigration, aspect, european	0.39	19

Table 1: List of top 20 dynamic topics, ranked by their TC-W2V topic coherence. For each dynamic topic, we report a manually-assigned short label, the top 10 terms, coherence, and frequency (*i.e.* number of windows in which it appeared).

terms). The presence of these higher-level associations between topics provide semantic validity for the results presented, where topics that one might expect to be related are found to be correlated with respect to rows in their NMF factor $\mathbf{H}$ (*i.e.* similar terms appear in the set of topic descriptors (words) that define them as topics).

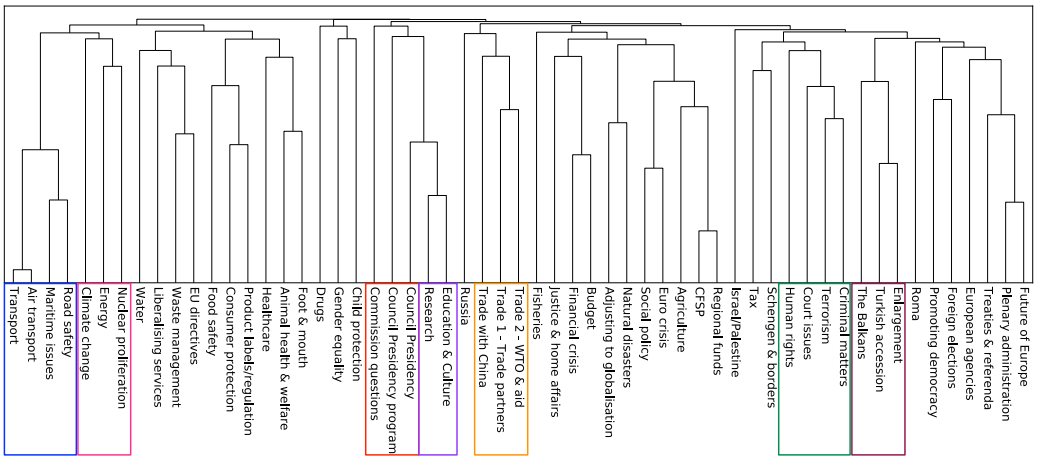


Figure 3: Dendrogram for average linkage hierarchical agglomerative clustering of 57 dynamic topics.

External Validation

The data analysis task performed in this paper is inherently unsupervised, in the sense that our corpus does not contain any annotated tags or labels indicating the nature of the content of speeches. Therefore, to assess the extent to which the dynamic topics identified correspond to EU policy areas, and thus provide evidence of construct validity, we compare the 57 dynamic topics to an existing taxonomy of subjects used by Europarl to classify legislative procedures. The taxonomy retrieved from the EP website has several different levels, ranging from broad top-level subjects (*e.g.* '3 Community policies'), to highly-specific low-level subjects (*e.g.* '3.10.06.05 Textile plants, cotton'). We compare our results to the second level of the taxonomy, containing 48 subjects (*e.g.* '3.10 Agricultural policy and economies', '3.20 Transport policy in general'). For each subject code, we create a "subject document" consisting of the description of the subject and all lower-level subjects within that branch of the taxonomy. We then identify the most similar dynamic topic by comparing the top 10 terms for that topic with subject documents, based on cosine similarity.

Table 2 shows the best matching subjects and topics identified using this approach. To give a couple of examples, the topic hand-coded as relating to 'Tax' from our topic model was correctly matched with the Europarl subject code '2.70 Taxation' broadly defined at level-2 of the taxonomy, and with '2.70.01 Direct taxation' and '2.70.02 Indirect taxation' defined separately at level-3 of the taxonomy. When looking at the topic manually labeled as relating to 'Drugs', cosine similarity matches this with the level-2 subject '4.20 Public health', which has a level-3 sub-category relating to '4.20.04 Pharmaceutical products and industry'.

Subject	Matched Topic: Top 10 Terms	Sim.
1.10 Fundamental Rights In The Union	rights, human, fundamental, freedom, democracy, law, charter, resolution, union, violation	0.66
4.40 Education, Vocational Training & Youth	programme, education, culture, language, cultural, youth, sport, learning, young, training	0.63
5.20 Monetary Union	euro, economic, growth, stability, pact, bank, policy, monetary, economy, ecb	0.62
4.70 Regional Policy	policy, region, cohesion, development, regional, strategy, structural, fund, economic, area	0.62
3.50 Research & Technological Development	research, programme, innovation, framework, funding, industry, technology, development, cell, institute	0.57
3.60 Energy Policy	energy, gas, renewable, efficiency, supply, source, electricity, market, target, project	0.53
6.10 Common Foreign & Security Policy	security, policy, defence, common, foreign, military, nato, immigration, aspect, european	0.52
3.20 Transport Policy in General	transport, railway, rail, passenger, road, network, freight, system, train, infrastructure	0.51
4.60 Consumers' Protection in General	product, medicinal, medicine, tobacco, labelling, safety, consumer, regulation, organic, advertising	0.50
3.70 Environmental Policy	waste, recycling, directive, packaging, management, environment, electronic, fuel, environmental, radioactive	0.50

Table 2: Top 10 legislative procedure subjects with corresponding matching dynamic topics, ranked by cosine similarity of the match.

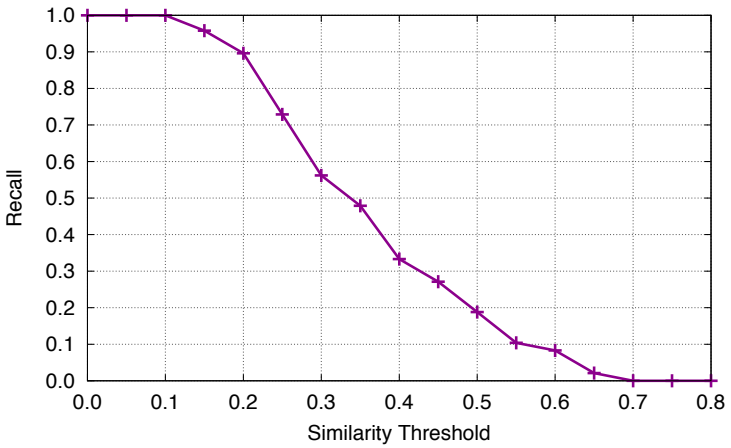


Figure 4: Recall plot for EP taxonomy subjects relative to dynamic topics, for increasing thresholds for cosine similarity.

When taken in the context of the matches shown in Table 2, this indicates that our dynamic topics provide good coverage of the policy areas that might be expected to feature during EP debates, and thus increases our confidence in the construct validity of the model.

Appendix C: Dynamic Comparison

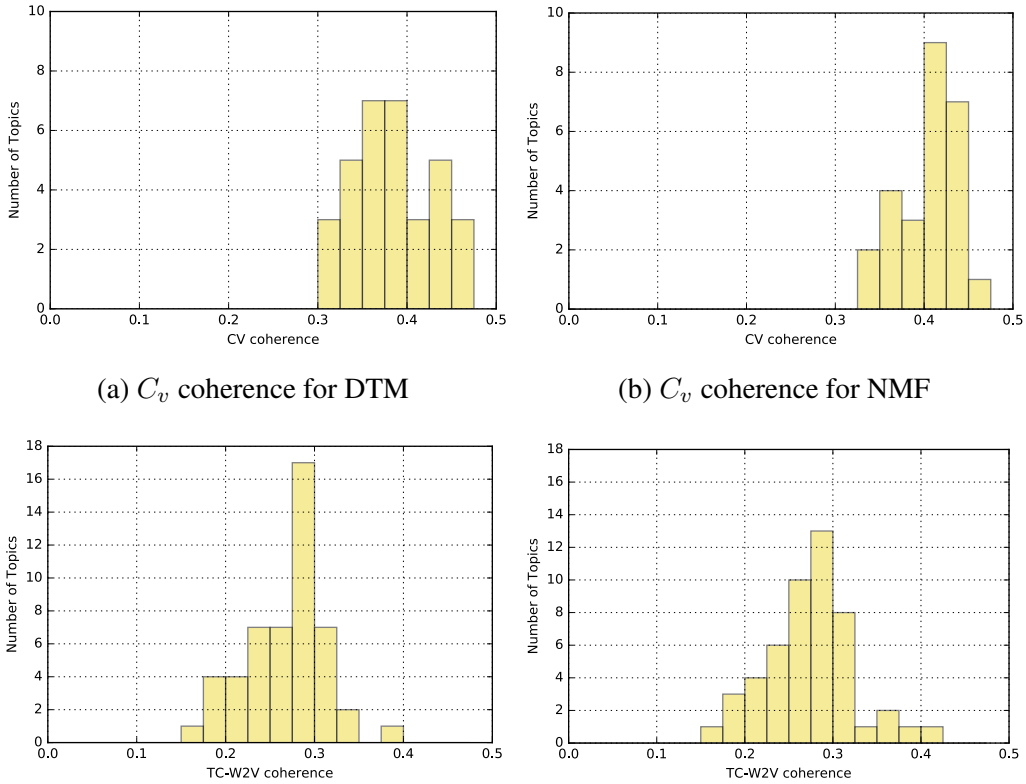
As an additional comparison, we also examined the application of the probabilistic Dynamic Topic Modeling (DTM) algorithm proposed by Blei and Lafferty (2006) to the corpus of parliamentary speeches. For the purpose of comparison, we apply both our proposed NMF-based approach and DTM for a fixed number of $k = 50$ dynamic topics to the entire corpus. In the case of NMF, we generate the first layer of window topic models as described in Section 6.1 of the paper. In the case of DTM, we use the original C++ implementation[1] and apply the algorithm using the default parameters recommended by the authors, using the same time window division as NMF.

When we compare the overall results, the two approaches were in broad agreement, particularly in relation to the identification of dynamic topics relating to general policy areas, such as security, agriculture, transport, and fisheries. To quantitatively compare the outputs, we assessed the coherence of the dynamic topics using the TC-W2V and C_v measures described in Section 5 of the paper, again using the top 10 terms to describe each topic. In the case of the C_v topic coherence measure, the NMF approach had a higher median coherence of 0.458 versus 0.424 for DTM. The NMF-based approach also yielded a marginally higher median TC-W2V coherence of 0.277 versus 0.276. The distribution of values for all 50 dynamic topics are shown in Fig. 5.

However, when we examine the actual window topics produced by each method, the results are quite different. Since the dynamic topics generated by DTM are built sequentially, the top terms reported at each time window are relatively stable. In contrast, with the NMF-based approach, each time window topic model is produced independently based only on the data present in that window. As a result, the top terms for each topic are far more indicative of the trends related to that topic at a given point in time. Table 3 shows a representative example, corresponding to the dynamic topics related to climate change found by both methods, when broken down to their window topics across five quarterly time windows. We see that the top 10 terms for the NMF-based topics are far more diverse, reflecting the changing nature of discussion items around climate change in the European Parliament, such as the Cancun Agreements reached on at the 2010 United Nations Climate Change Conference in Mexico.

To examine this difference quantitatively, for both topic modeling methods we look at the agreement between the top ranked terms consecutive pairs of window topics in each of the $k = 50$ dynamic topics. We quantify the agreement between two term rankings using the Jaccard coefficient, which is the size of the intersection of the term sets divided by the size of their union. A score of 1 indicates that

[1] `http://www.cs.princeton.edu/~blei/topicmodeling.html`



(a) C_v coherence for DTM

(b) C_v coherence for NMF

(c) TC-W2V coherence for DTM

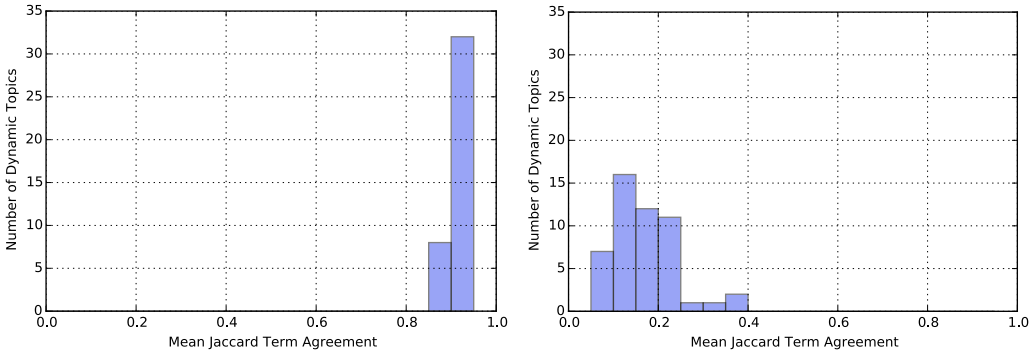
(d) TC-W2V coherence for NMF

Figure 5: Distributions of coherence scores for $k = 50$ dynamic topics, comparing the probabilistic and the NMF-based dynamic topic modeling methods.

Window	**NMF Window Topic**	**DTM Window Topic**
2008-Q4	energy, climate, emission, package, change, renewable, target, industry, carbon, gas	energy, climate, change, gas, european, emission, package, supply, efficiency, renewable
2009-Q1	climate, change, future, emission, integrated, water, policy, target, industrial, global	energy, climate, change, gas, european, emission, efficiency, supply, package, renewable
2009-Q4	climate, change, copenhagen, developing, emission, conference, summit, agreement, global, energy	energy, climate, change, copenhagen, european, gas, emission, efficiency, supply, carbon
2010-Q1	climate, copenhagen, change, summit, emission, international, mexico, conference, global, world	energy, climate, change, european, copenhagen, gas, emission, efficiency, supply, carbon
2010-Q4	climate, trade, change, cancun, conference, international, agreement, emission, environmental, global	energy, climate, change, european, gas, efficiency, emission, supply, target, source

Table 3: Example of window topics associated with a dynamic topic related to climate change, produced by both the NMF-based approach and DTM on the same time window datasets.

the term sets are identical (not considering rank order), while a score of 0 indicates



(a) Jaccard Term Agreement for DTM (b) Jaccard Term Agreement for NMF

Figure 6: Distributions of Jaccard term agreement scores $k = 50$ dynamic topics, for the probabilistic and the NMF-based dynamic topic modeling methods.

that the sets share no terms in common. For each dynamic topic generated by the two methods, we calculate the mean agreement between the consecutive window topics form which it is composed.

Fig. 5 shows the distribution of Jaccard agreement scores for the dynamic topics produced by both methods. We see a stark difference between the extent to which the terms associated with each topic change over time – the overall mean Jaccard score across all dynamic topics for the NMF-based approach is 0.166, reflecting the fact that the top terms change frequently over time. In contrast, the overall mean score is 0.921 for the probabilistic approach indicates that the top terms often remain fixed and do not change frequently over time. Therefore, although the descriptors for the overall dynamic topics are relatively similar in terms of their coherence, when we wish to explore the time windows from which they are assembled, the NMF-based approach yields topics that more closely reflect the parliamentary discussions during each window, thereby supporting the interpretation of the topics.

References

Blei, D. M. and Lafferty, J. D. (2006) 'Dynamic topic models', in 'Proc. 23rd International Conference on Machine Learning', pp. 113–120.

Greene, D., Cagney, G., Krogan, N. and Cunningham, P. (2008) 'Ensemble Non-negative Matrix Factorization Methods for Clustering Protein-Protein Interactions'. *Bioinformatics*, Vol. 24, No. 15, pp. 1722–1728.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H. and Radev, D. R. (2010) 'How to analyze political attention with minimal assumptions and costs'. *American J. Political Science*, Vol. 54, No. 1, pp. 209–228.