Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis

Derek Greene
School of Computer Science & Informatics
University College Dublin, Ireland
derek.greene@ucd.ie

James P. Cross
School of Politics & International Relations
University College Dublin, Ireland
james.cross@ucd.ie

ABSTRACT

This study analyzes political interactions in the European Parliament (EP) by considering how the political agenda of the plenary sessions has evolved over time and the manner in which Members of the European Parliament (MEPs) have reacted to external and internal stimuli when making Parliamentary speeches. It does so by considering the context in which speeches are made, and the content of those speeches. To detect latent themes in legislative speeches over time, speech content is analyzed using a new dynamic topic modeling method, based on two layers of matrix factorization. This method is applied to a new corpus of all English language legislative speeches in the EP plenary from the period 1999-2014. Our findings suggest that the political agenda of the EP has evolved significantly over time, is impacted upon by the committee structure of the Parliament, and reacts to exogenous events such as EU Treaty referenda and the emergence of the Euro-crisis have a significant impact on what is being discussed in Parliament.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

Keywords

Topic modeling, Text mining, Political speech, EU politics

1. INTRODUCTION

The plenary sessions of the European Parliament (EP) are one of the most important arenas in which European representatives can air questions, express criticisms and take policy positions to influence EU politics. Indeed the plenary of the Parliament represents the closest that the European Union (EU) gets to engaging in the core democratic process of publicly-aired democratic debate. As a result, understanding how Members of the European Parliament (MEPs) express themselves in plenary, and investigating how the political discussions evolve and respond to internal and external stimuli is a fundamentally important undertaking.

In recent years, there has been a concurrent explosion of online records detailing the content of MEP speeches, and the development of data mining techniques capable of extracting latent patterns in content across sets of these speeches. This allows us for the first time to investigate the plenary agenda of the Parliament in a holistic and rigorous manner. One approach to tracking the political attention of political figures has been to apply topic modeling algorithms to large corpora of political texts, such as parliamentary speeches of the U.S. Senate [18]. These algorithms seek to distill the latent thematic patterns in a corpus of speeches [3], and can be used to improve the transparency of the political process by providing a macro-level overview of the activities and agendas of politicians in a time- and resource-efficient manner. This type of overview would otherwise be unavailable due to the time and resource costs associated with manually hand-coding such a large-scale corpus.

This paper takes up the challenge of extracting latent thematic patterns in political speeches by developing a suitable dynamic topic modeling method to investigate how the plenary agenda of the EP has changed over three parliamentary terms (1999–2014), based on the analysis of a corpus of 210,247 speeches from 1,735 MEPs across the 28 EU member states. The method described in Section 3 involves applying two layers of Non-negative Matrix Factorization (NMF) topic modeling [12]. Firstly, the corpus of speeches is divided into distinct segments or *time windows*, on which low-level *window topics* are identified by applying NMF. Secondly, the topics from each window are represented as a combined matrix of "topic documents". By applying NMF to this new representation, we can identify high-level *dynamic topics* which potentially span many time windows. This process allows us to explore parliamentary activity both at a granular level and over multiple parliamentary terms. In addition, we can relate these dynamic topics to the original speakers, allowing us to identify the topics to which individual MEPs are paying most attention.

Applying our proposed topic modeling methodology reveals the breadth of policy areas covered by the EP, and the results presented later in Section 5 indicate that the political agenda of the Parliament has evolved significantly across the three parliamentary terms considered here. By examining a number of topic case studies, ranging from the Euro-crisis to EU treaty changes, we can identify the relationship between the evolution of these dynamic topics and the exogenous events driving them. By using external data sources, we can also confirm the semantic and construct validity of these topics. In order to explain some of the patterns in speech making we observe, we conclude the paper with an exploration of the determinants of MEP speech-making behavior on the topics detected by our topic model. To provide access to the results of the project to interested parties, we make a browsable version available online[1]. This website provides a greater level of transparency into the activities of the EP as a functioning democratic institution.

[1] `http://erdos.ucd.ie/europarl`

WebSci '15, June 28 - July 01, 2015, Oxford, United Kingdom.

http://dx.doi.org/10.1145/2786451.2786464.

2. RELATED WORK

2.1 European Parliament

The most prominent forms of MEP behavior that have been examined in the existing literature include the expression of policy positions through speeches and written submissions, and voting in plenary. The formal committee structure of the Parliament provides strategic advantages to certain MEPs by providing committee members with privileged access to information, and opportunity to shape the Parliament's negotiation stance. This has led MEPs to self-select into committees dealing with issues that they find salient in order to affect outcomes in those policy areas [5, 26].

Committee chairs hold important administrative powers to set the committee agenda and the topics for debate at committee meetings. Rapporteurs are tasked with preparing reports about committee activities, and represent a medium for disseminating information about committee activities to the broader plenary [1]. Rapporteurs thus plays a central role in shaping the image of committee activities available to committee outsiders. Outside of committees, strict institutional rules also govern the allocation of speaking time during the Parliament's plenary sessions, and structure the ability of MEPs to intercede during negotiations [10, 16]. The total amount of speaking time for any particular issue is limited and divided between time reserved for actors with formal duties in plenary such as rapporteurs, and time proportionally divided between party groups based upon their overall share of MEPs elected. Limits on speaking time can lead to competition between MEPs, and party group leaders allocate the scare resource of speaking time between competing demands from rank and file MEPs for maximum impact.

Due to the limits in the total amount of speaking time available, MEPs can also submit written questions and statements that are appended to the plenary records. These provide extra opportunity for MEPs to state their positions outside the time limits imposed on oral questioning during plenary debates. These written questions have been found to be the most popular avenue used by MEPs to interact with the Commission directly [19], and provide the opportunity for 'fire-alarm oversight' of national governments guilty of implementation failures of EU law [11]. MEPs enjoy more discretion over their ability to submit written submissions than they do over oral speaking time.

In terms of the content of legislative speeches in the Parliament, it has been shown that speeches reflect latent ideological conflict between MEPs, with both left-right and pro-/anti-EU integration dimensions of conflict having been detected [22]. Using text analysis techniques based upon word-frequency distributions, these authors were able to demonstrate the correspondence between the content of legislative speeches and other measures of ideological positions found in the literature based upon roll-call votes and expert surveys.

2.2 Topic Models

In the field of data analytics advanced topic modeling algorithms that go beyond word-frequency distributions have recently been applied to large-scale text collections. Considerable research on topic modeling has focused on the use of probabilistic methods such as variants of Latent Dirichlet Allocation (LDA) [23]. Authors have subsequently developed analogous probabilistic approaches for tracking the evolution of topics over time in a sequentially-organized corpus of documents, such as the dynamic topic model (DTM) of Blei and Lafferty [2]. Alternative algorithms, such as Non-negative Matrix Factorization (NMF) [12], have also been effective in discovering the underlying topics in text corpora [15, 25]. Saha & Sindhwani [20] proposed an online learning framework for employing NMF to extract topics from streaming social media content, by dividing the streams into short sliding time windows so as discover topics that are smoothly evolving over time.

As well as analyzing temporal data, recent work in this area has focused on important practical issues, including automating parameter selection (*e.g.* how many topics are appropriate for our corpus?) and assessing *topic coherence* (*i.e.* how meaningful are the topics generated by our algorithm?) [7, 15]. The latter corresponds closely to the concept of *semantic validity* introduced in [18] for assessing the reliability of topics found in text corpora. This concept covers both intra-topic validity (the extent to which a single topic is meaningful) and inter-topic validity (the extent to which different topics are related to one another in a meaningful way).

2.3 Topic Models Applied to Political Texts

Some topic modeling methods have been adopted in the political science literature to analyze political attention. In settings where politicians have limited time-resources to express their views, such as the plenary sessions in parliaments, politicians must decide what topics to address. Analyzing such speeches can thus provide insight into the political priorities of the politician under consideration. Single membership topic models, that assume each speech relates to one topic, have successfully been applied to plenary speeches made in the 105th to the 108th U.S. Senate in order to trace political attention of the Senators within this context over time [18]. This study found that a rich and meaningful political agenda emerged from the collected speeches, where topics evolved significantly over time in response to both internal and external stimuli.

Bayesian hierarchical topic models have also been used to capture the political priorities of Members of Congress as found in their official press releases [9]. This study shows that the press releases are also responsive to external stimuli such as upcoming votes in Congress or events external to Congress such as the anniversary of September 11th. Press release topics are also geographically structured with Members of Congress from rural farming communities more likely to pay attention to agricultural issues than those from urban communities for instance. The introduction of these methods to the study of political attention has allowed researchers to consider larger and more complete datasets of political activity across longer time periods than has previously been possible. The results unveil latent patterns in political attention that are difficult and time-consuming to capture using more traditional methodological approaches, such as expert surveys and hand-coding of texts. Applying them to study the political agenda of the European Parliament is the aim of this paper.

3. METHODS

In this section we describe a two-layer strategy for applying topic modeling in a non-negative matrix factorization framework to a timestamped corpus of political speeches. Firstly, in Section 3.1 we describe the application of NMF topic modeling to a single set of speeches from a fixed time period. Secondly, in Section 3.2 we propose a new approach for combining the outputs of topic modeling from successive time periods to detect a set of *dynamic topics* that span part or all of the duration of the corpus.

3.1 Topic Modeling Speeches

While work on topic models often involves the use of LDA, NMF can also be applied to textual data to reveal topical structures [25]. The ability of NMF to apply TF-IDF weighting to the data prior to topic modeling has shown to be advantageous in producing diverse but semantically coherent topics which are less likely to be represented by the same high frequency terms. This makes NMF suitable when the task is to identify both broad, high-level groups

of documents and niche topics with specialized vocabularies [15].

Given a corpus of n speeches, we first construct a speech-term frequency matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$, where m is the number unique terms present across all speeches (*i.e.* the corpus vocabulary). Applying NMF to $\mathbf{A}$ results in a reduced rank-k approximation in the form of the product of two non-negative factors $\mathbf{A} \approx \mathbf{WH}$, where the objective is to minimize the reconstruction error between $\mathbf{A}$ and $\mathbf{WH}$. The rows of the factor $\mathbf{H} \in \mathbb{R}^{k \times m}$ can be interpreted as k topics, defined by non-negative weights for each of the m terms in the corpus vocabulary. Ordering each row provides a topic descriptor, in the form of a ranking of the terms relative to corresponding topic. The columns in the matrix $\mathbf{W} \in \mathbb{R}^{n \times k}$ provide membership weights for all n speeches with respect to the k topics.

In our experiments we use the fast alternating least squares variant of NMF introduced in [13]. NMF algorithms are often initialized with random factors. However, this can lead to unstable results, where the algorithm converges to a variety of different local minima of poor quality, depending on the random initialization. To ensure a deterministic output and to improve the quality of the resulting topics, we generate initial factors using the Non-negative Double Singular Value Decomposition (NNDSVD) approach [4].

A key parameter selection decision in topic modeling pertains to the number of topics k. Choosing too few topics will produce results that are overly broad, while choosing too many will lead to many small, highly-similar topics. One general strategy proposed in the literature has been to compare the *topic coherence* of topic models generated for different values of k [7]. A range of such coherence measures exist in the literature, although many of these are specific to LDA. Recently, O'Callaghan *et al.* [15] proposed a general measure, TC-W2V, which evaluates the relatedness of a set of top terms describing a topic, based on the similarity of their representations in a *word2vec* distributional semantic space [14]. Specifically, the coherence of a topic t_h represented by its t top ranked terms is given by the mean pairwise cosine similarity between all relevant term vectors in the *word2vec* space:

$$\text{coh}(t_h) = \frac{1}{\binom{t}{2}} \sum_{j=2}^{t} \sum_{i=1}^{j-1} cos(wv_i, wv_j) \quad (1)$$

An overall score for a topic model T consisting of k topics is given by the mean of the individual topic coherence scores:

$$\text{coh}(T) = \frac{1}{k} \sum_{h=1}^{k} \text{coh}(t_h) \quad (2)$$

An appropriate value for k can be identified by examining a plot of the mean TC-W2V coherence scores for a fixed range $[k_{min}, k_{max}]$ and selecting a value corresponding to the maximum coherence. An example is shown in Fig. 1, where the plot of mean coherence scores suggests a value $k = 19$ from a candidate range $[10, 25]$.

Parliamentary speeches will often be short and concise. In the case of the EP, speeches are often limited to 1-2 minutes in duration. As such, we would expect each speech to be primarily related to a single topic. This is consistent with the observations made by Quinn *et al.* [18] when analyzing speeches from the U.S. Congress. Here we produce a single membership topic model (*i.e.* a disjoint clustering of individual speeches in relation to topics) by selecting the maximum membership weight for each row in the factor $\mathbf{W}$.

3.2 Dynamic Topic Modeling

When applying clustering to temporal data, authors have often proposed dividing the data into *time windows* of fixed duration [24]. In the case of streaming data, such as content originating from social media platforms, this involves artificially transforming

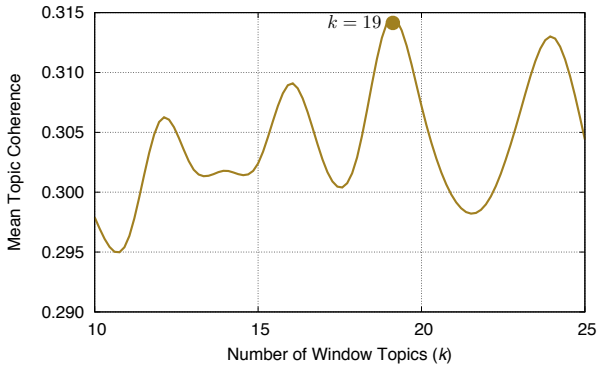


Figure 1: Plot of mean TC-W2V topic coherence scores for different values for the number window topics k, generated on a time window of European Parliament speeches from 2005-Q1.

the continuous streams into sliding windows. However, in the case of political speeches transcribed from distinct plenary sessions, the data is naturally divided into segments. While some aspects of the agenda will remain common between successive sessions, in other cases the focus of debates will change considerably between sessions. Consequently, online learning approaches which use sliding windows and assume a smooth evolution in topics over time, such as proposed in [20], may be unsuitable.

Following [24], we divide the full time-stamped corpus of parliamentary speeches into τ disjoint time windows $\{W_1, \ldots, W_\tau\}$ of equal length. The rationale for the use of time windows as opposed to processing the full corpus in batch is two-fold: 1) we are interested in identifying the agenda of the parliament at individual time points as well as over all time; 2) short-lived topics, appearing only in a small number of time windows, may be obscured by only analyzing the corpus in its entirety. At each time window W_i, we apply NMF with parameter selection based on Eqn. 2 to the transcriptions of all speeches delivered during that window, yielding a *window topic model* T_i containing k_i *window topics*. This process produces a set of successive window topic models $\{T_1, \ldots, T_\tau\}$, which represents the output of the first layer in our proposed methodology.

From the window topic models we construct a new condensed representation of the original corpus, by viewing the rows of each factor $\mathbf{H}_i$ coming from each window topic model as "topic documents". Each topic document naturally contains non-negative weights indicating the descriptive terms for that window topic. We expect that window topics from different windows which share a common theme will have similar topic documents. Specifically, we construct a condensed topic-term matrix $\mathbf{B}$ as follows:

1. Start with an empty matrix $\mathbf{B}$.
2. For each window topic model T_i:
   (a) For each window topic within T_i, select the t top ranked terms from the corresponding row vector of the associated NMF factor $\mathbf{H}$, set all weights for all other terms in that vector to 0. Add the vector as a new row in $\mathbf{B}$.
3. Once vectors from all topic models have been stacked in this way, remove any columns with only zero values (*i.e.* terms from the original corpus which did not ever appear in the t top ranked terms for any window topics).

The matrix $\mathbf{B}$ has size $n' \times m'$, where $n' = \sum_{i=1}^{\tau} k_i$ is the total number of "topic documents" and $m' << m$ is is the subset of relevant terms remaining after Step 3. The use of only the top t terms in each topic document allows us to implicitly incorporate feature selection into the process. The result is that we include those terms that were highly-descriptive in each time window, while excluding

Rank	2008-Q4	2009-Q1	2009-Q4	2010-Q1
1	energy	climate	climate	climate
2	climate	change	change	copenhagen
3	emission	future	copenhagen	change
4	package	emission	developing	summit
5	change	integrated	emission	emission
6	renewable	water	conference	international
7	target	policy	summit	mexico
8	industry	target	agreement	conference
9	carbon	industrial	global	global
10	gas	global	energy	world

Table 1: Example of 4 window topics, described by lists of top 10 terms, which have been grouped together in a single dynamic topic related to climate change.

those terms that never featured prominently in any window topic. This reduces the computational cost for the second factorization procedure described below.

Having constructed $\mathbf{B}$, we now apply a second layer of NMF topic modeling to the matrix to identify k' *dynamic topics* which potentially span multiple time windows. The process is the same as that outlined previously in Section 3.1. Here the TC-W2V coherence measure is used to detect number of dynamic topics k'. The resulting factors can be interpreted as follows: the top ranked terms in each row of $\mathbf{H}$ provide a description of the dynamic topics; the values in the columns of $\mathbf{W}$ indicate to what extent each window topic is related to each dynamic topic.

We track the evolution of these topics over time as follows. Firstly, we assign each window topic to the dynamic topic for which it has the maximum weight, based on the values in each row in the factor $\mathbf{W}$. We define the temporal *frequency* of a dynamic topic as the number of distinct time windows in which that dynamic topic appears. The set of all speeches related to this dynamic topic across the entire corpus corresponds to the union of the speeches assigned to the individual time window topics which are in turn assigned to the dynamic topic. To summarize, the key outputs of the two-layer topic modeling process are as follows:

1. A set of τ topic models, one per time window, each containing k_i *window topics*. These are described using their top t terms and the set of all associated speeches.
2. A set of k' *dynamic topics*, each with an associated set of window topics. These are described using their top t terms and set of all associated speeches.

Table 1 shows a partial example of a dynamic topic. We observe that, for the four window topics, there is a common theme pertaining to climate change. While the variation across the term lists reflects the evolution of this dynamic topic over the corresponding time period (2008-Q4 to 2010-Q1), the considerable number of terms shared between the lists underlines its semantic validity.

4. DATA

During August 2014 we retrieved all plenary speeches available on Europarl, the official website of the European Parliament[2], corresponding to parliamentary activities of MEPs during the 5th – 7th terms of the EP. This resulted in 269,696 unique speeches in 24 languages. While we considered the use of either multi-lingual topic modeling or automated translation of documents, issues with the accuracy and reliability of both strategies lead us to focus on English language speeches in plenary – either from native speakers or translated – which make up the majority of the speeches available on Europarl. A corpus of 210,247 English language speeches was identified in total, representing 77.95% of the original collection. In terms of coverage of speeches from MEPs from the member states, this ranged from 100% for the United Kingdom, through 87% for Germany, down to 66.2% for Romania. However, the most recent state to accede to the EU, Croatia, represents an outlier in the sense that only 2.6% of speeches were available in English at the time of retrieval due to EP speech translation issues.

We subsequently divided the corpus into 60 quarterly time windows, from 1999-Q3 to 2014-Q2. We selected a quarter as the time window duration to allow for the identification of granular topics, while also ensuring there there existed a sufficient number of speeches in each time window to perform meaningful topic modeling. In particular, we wished to avoid empty time windows occurring due to the summer recess of the EP. For each time window W_i we construct a speech-term matrix $\mathbf{A}_i$ as follows:

1. Select all speech transcriptions from window W_i, and remove all header and footer lines.
2. Find all unigram tokens in each speech, through standard case conversion, tokenization, and lemmitization.
3. Remove short tokens with < 3 characters, and tokens corresponding to generic stop words (*e.g.* "are", "the"), parliamentary-specific stop words (*e.g.* "adjourn", "comment") and names of politicians as listed on the EP website.
4. Remove tokens occurring in < 5 speeches.
5. Construct $\mathbf{A}_t$, based on the remaining tokens. Apply standard TF-IDF term weighting and document length normalization.

The resulting time window data sets range in size from 679 speeches in 2004-Q3 to 9,151 speeches in 2011-Q4, with an average of 4,811 terms per data set.

5. EXPERIMENTAL RESULTS

5.1 Experimental Setup

After pre-processing the data, to identify window topics we applied NMF with parameter selection as described in Section 3.1. Given the relatively specialized vocabulary used in EP debates, when building the *word2vec* space for parameter selection, as our background corpus we used the complete set of English language speeches. We used the same *word2vec* settings and number of top terms per topic ($t = 10$) as described in [15]. At each time window, we generated models containing $k \in [10, 25]$ window topics, selecting the value k that maximized mean TC-W2V coherence. The resulting median number of topics per window was 15.5. The illustration of the number of topics per window in Fig. 2 shows that there is considerable variation in the number of topics detected for each window, which does not correlate with the number of speeches per quarter (Pearson correlation 0.006).

The process above yielded 1,017 window topics across the 60 time window. We subsequently applied dynamic topic modeling as described in Section 3.2. For the number of terms t representing each window topic, we experimented with values from 10 to the entire number of terms present in a time window. However, values $t > 20$ did not result in significantly different dynamic topics. Therefore, to minimize the dimensionality of the data, we selected $t = 20$. This yielded a matrix of 1,017 window topics represented by 2,710 distinct terms. We applied parameter selection based on TC-W2V coherence to select an overall number of dynamic topics k' from a range $k' \in [25, 90]$. The resulting plot (see Fig. 3) indicated a maximal value at $k' = 57$, although a number of close peaks exist in the range [62,80]. This corresponds to our manual inspections of the results, where the topic models for these values of k' appeared to be highly similar, with minor variations corresponding to merges or splits of strongly-related topics.

[2] http://europarl.europa.eu

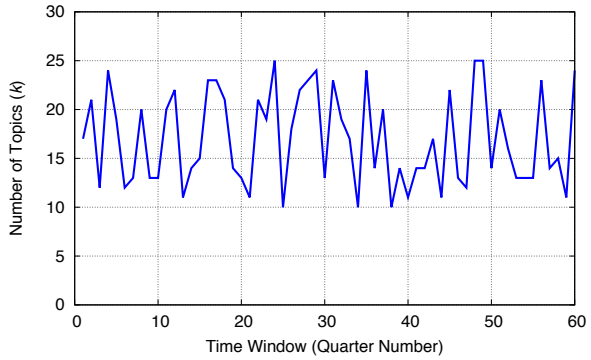


Figure 2: Number of window topics identified per time window, from 1999-Q3 (#1) to 2014-Q2 (#60).

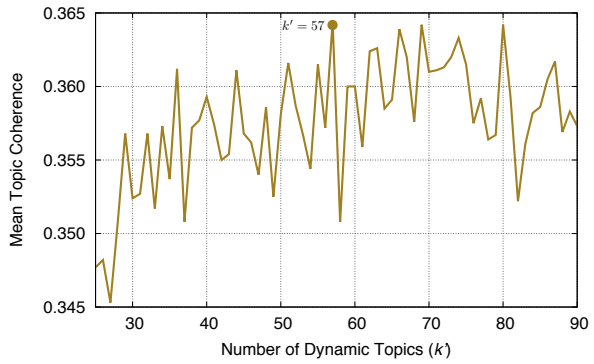


Figure 3: Plot of mean TC-W2V topic coherence scores for different values for the number dynamic topics of k', across a candidate range $[25, 90]$.

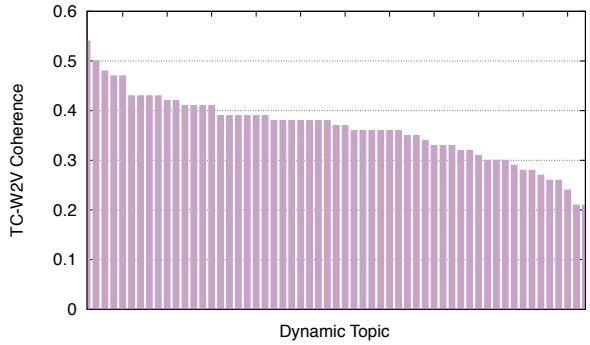


Figure 4: Distribution of TC-W2V topic coherence values for 57 dynamic topics, based on top 10 terms.

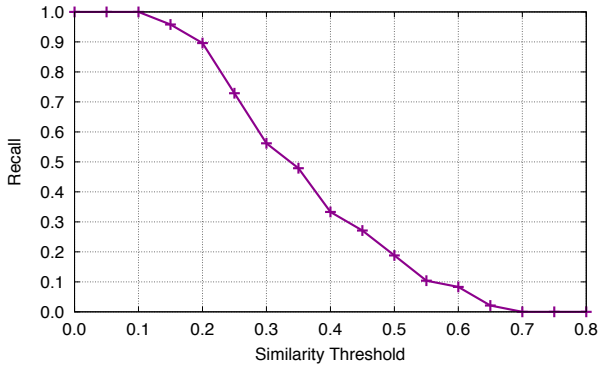


Figure 5: Recall plot for EP taxonomy subjects relative to dynamic topics, for increasing thresholds for cosine similarity.

5.2 Dynamic Topic Validation

The 57 topics identified in our experiments are diverse, both in terms of their thematic content and temporal signatures. Table 2 lists the top 20 dynamic topics in the data, ranked with respect to their TC-W2V topic coherence scores. We report the temporal frequency of the topics, together with a manually-assigned short label for discussion purposes[3]. The frequency of dynamic topics ranged from 11 which appeared in < 10 time windows, to a broad 'Plenary administration' topic which appeared in 57 out of 60 windows.

In general, we observed two distinct categories of dynamic topics. The first reflects the day-to-day politics of EU in terms of legislating and debating issues related to the core EU competencies (*e.g.* 'Energy', 'Agriculture'), while the other reflects unanticipated exogenous shocks and MEPs reactions to these events (*e.g.* Euro-crisis, September 11th attacks). These two categories exhibit differing temporal signatures. For instance, we see a considerable difference between the broad topic on fisheries policy (Fig. 7(a)), when compared to the two topics arising from the events during the financial crisis and subsequent Euro-crisis as shown in Figures 7(b) and 7(c) respectively. This distinction between dynamic topic types reflects two different forms of political process in the Parliament.

To examine the intra-topic semantic validity of these dynamic topics, Fig. 4 illustrates the distribution of TC-W2V coherence values for all dynamic topics, when evaluated in the *word2vec* space built from the complete speech corpus. As evidence by the ranking in Table 2, 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 (-1.0), suggesting a high level of semantic validity across the board.

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. Following the approach described in [8], 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. 6. Using the interpretation provided in [18], the lower the height at which any two topics are connected in the dendrogram, the more similar their corresponding term usage patterns in EP sessions. We observe a number of higher-level groupings of interest, which are highlighted in Fig. 6. These include groups related to transport in general, energy concerns, interactions with other institutions, education and research, trade relations, and EU enlargement. The presence of these higher-level associations between topics provide semantic validity for the results presented, where topics one might expect to be related are found to be correlated with respect to their rows in the NMF factor $\mathbf{H}$ (*i.e.* they share similar terms in their topic descriptors).

To externally quantify the extent to which the identified dynamic topics correspond to policy areas in which the EU has competencies, and thus provide evidence of construct validity, we compare

[3]Full details of all window topics and dynamic topics are available at `http://erdos.ucd.ie/europarl`

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 and 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 and aid	trade, wto, world, development, developing, international, negotiation, aid, free, relation	0.39	19
47	Product labelling and 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 2: 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 time windows in which it appeared).

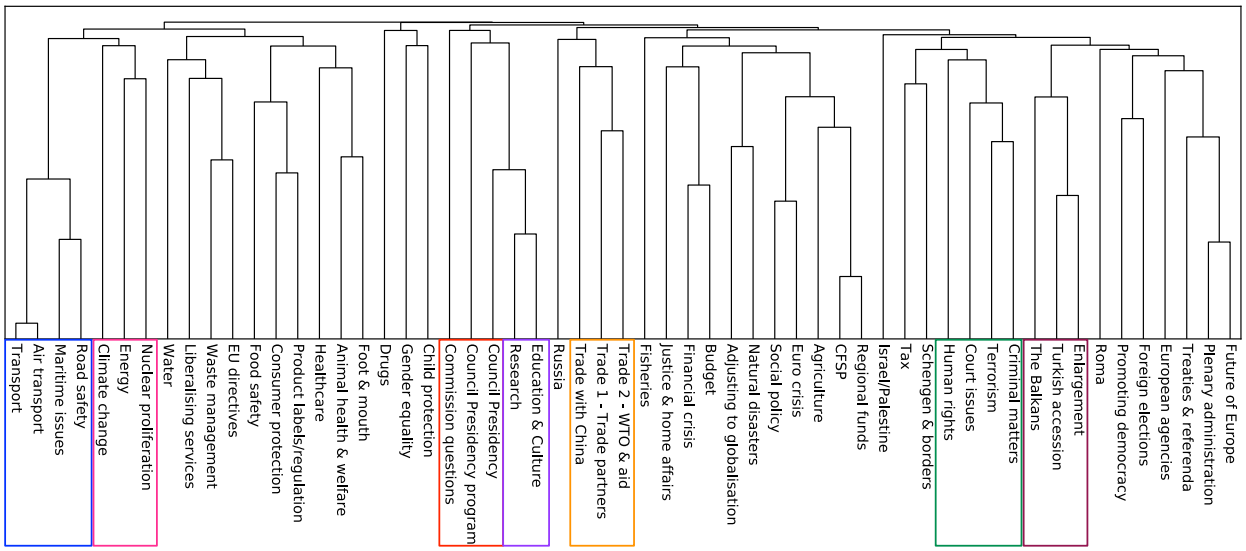


Figure 6: Dendrogram for average linkage hierarchical agglomerative clustering of 57 dynamic topics.

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

the 57 dynamic topics to an existing taxonomy of subjects, which is used by Europarl to classify legislative procedures. The taxonomy as retrieved from the site 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 each subject document, based on cosine similarity. Table 3 shows the best matching subjects and topics identified using this approach. Fig. 5 shows the *recall* of all 48 subjects, for different threshold levels of cosine similarity. For instance, at a threshold of 0.25, suitably matching dynamic topics for 72.9% of subjects are identified. To give a couple of examples, the topic hand-coded as relating to 'Tax' from our topic model was

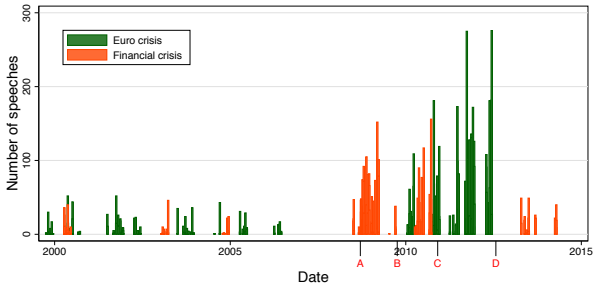
correctly matched with the Europarl subject code '2.70 Taxation' broadly defined at level-2 of the taxonomy, and '2.70.01 Direct taxation' and '2.70.02 Indirect taxation' separately at level-3 of the taxonomy. When looking at the topic manually labeled as relating to 'Drugs', cosine similarity matches 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'. When taken in the context of the matches shown in Table 3, 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 method.

5.3 Case Studies

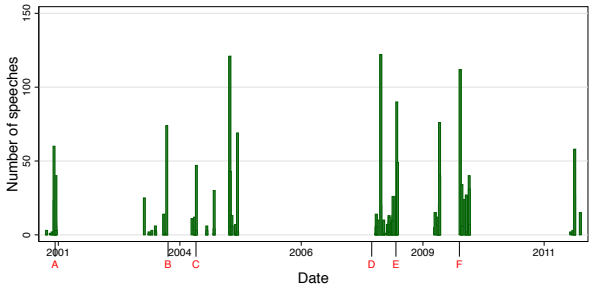
In order to further investigate the construct validity of our topics, we focus on three specific examples that demonstrate how our modeling strategy captures variation in MEP attention to a topic over time, and how this attention is impacted upon by external stimuli.

Our first case study relates to MEP attention to the financial/Euro-crisis. The temporal distribution of speeches relating to this topic is illustrated in Fig. 7(a). This is an interesting case study, as the initial financial crisis peaked in 2008, and the Euro-crisis that followed has gone through a number of phases with major events in 2009, 2010 and 2012. As such, these events can be thought of as exogenous shocks that only garner MEP attention after they occur, and their exogenous nature provides a way to externally validity the dynamic topic modeling approach in use here. Fig. 7(a) demonstrates a number of distinct peaks in MEP speech making on both the financial crisis topic (in orange) and the Euro-crisis topic (in green). Attention to the financial crisis starts to rise in 2008-Q3 and initially peaks in 2008-Q4 (point A in Fig. 7(a)). This peak in activity corresponds to the date when the Lehman Brothers investment bank collapsed (15/9/2008). The other peaks in activity in Fig. 7(a) correspond to important events in the Euro-crisis. Point B corresponds to the revelations about under-reporting of Greek debt following the Greek parliamentary election in October 2010, Point C to the Irish bailout in November 2010, and Point D to Mario Draghi's statement that the ECB was "ready to do whatever it takes to preserve the euro" in the July 2012 respectively.

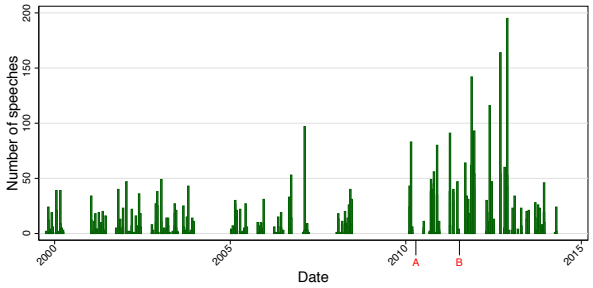
Our second case study relates to the process of EU treaty reform. This topic is of interest, because one would expect a large amount in variation in MEP attention to the topic over time, as Treaty revision and reform and the referenda that accompany them are rare event and should only garner MEP attention when such events occur. Fig. 7(b) shows MEP attention to the treaty change and referenda topic between 2000 and 2014 in terms of the number of speeches associated with this topic. Three distinct treaties were discussed and debated over this period. The first was the Nice treaty, which was agreed upon in 2001 and put to the vote in a referendum in Ireland in June 2001. The 'No' vote in that resulted from this referendum accounts for Point A in Fig. 7(b). The next set of treaty related events to occur were the negotiations and failed ratification of the Constitutional Treaty between 2003 and 2005. This process accounts for Point B in Fig. 7(b), that correspond to the Intergovernmental Conference negotiating the treaty text that begun in October 2003. In the end the Constitutional Treaty was rejected by the French and Dutch in referenda in May/June 2005. Point C indicates the date of the signing of the Enlargement treaty in May 2004. The Lisbon treaty was negotiated to replace the failed Constitutional treaty, and we observe a significant peak in MEP speeches directly relating to the Lisbon treaty when it was signed (Point D), and when the first Irish referendum failed to ratify the treaty in June 2008 (Point E). A similar peak in MEP speeches relating to treaty reform corresponds to the second Irish referendum that eventually



(a) "Financial & Euro crises" dynamic topics



(b) "Treaty changes & referenda" dynamic topic



(c) "Fisheries" dynamic topic

Figure 7: Time plots for three sample dynamic topics across all time windows, from 1999-Q3 (#1) to 2014-Q2 (#60). Dates on the x-axis correspond to the dates on which speeches were made at EP plenary sessions.

approved the Lisbon treaty in October 2009 (Point F).

Our third and final case study relates to fisheries policy. Fisheries is an interesting theme for the dynamic topic modeling approach to detect, because it is more associated with the day-to-day functioning of the EU as a regulator of the fisheries industry, when compared to more headline making policies and events like the economic crisis and treaty changes. Fig. 7(c) demonstrates the prevalence of the fisheries topic over time. As can be seen, MEPs are seen to pay a reasonably stable level of attention to fisheries in terms of the numbers of speeches being made between 2000 and 2010. This trend is interrupted in 2010, when an increase in MEP attention to the fisheries topic is observed. This can be explained by the fact that in 2009 the European Commission launched a public consultation on reforming EU fisheries policy, the results of which were presented to the Parliament and Council in April 2010. The launch of this working document corresponds to a increase in the number of MEP speeches related to the fisheries topic as detected by the dynamic topic model (Point A). The peak in MEP speech making relating to this topic (Point B) corresponds with Commissioner Maria Damanaki introducing a set of legislative proposals

designed to reform the common fisheries policy in a speech to the European Parliament in July 2011.

In general, the fact that the variation over time that we observe in MEP attention to these case study topics appears to be driven by exogenous events provides a form of construct validity for our topic modeling approach.

5.4 Explaining MEP Speech Counts

We now focus our attention on the 7th European Parliament which sat between 2009–2014. We focus on this term as a set of interesting covariates are available at the MEP level that can help us explain MEP contributions to a given topic. The dependent variable we seek to explain is the observed variation in the number of speeches each MEP makes on each of our identified dynamic topics. We employ a count-model framework suitable for analyzing count data [6]. The first issue to note with the count variable under consideration is that there is a large number of zeros. This is due to the fact that, for many topics, a considerable number of MEPs are recorded as making no speeches. This is likely due to the data-generating process in the topic model from which our dependent variable emerges. As described in Section 3.1, we apply a single membership topic modeling approach where each speech is associated with one topic. This assumption is generally unproblematic, given the short amount of time allowed for speeches and the concentrated nature of the messages MEPs seek to communicate in them. However, any speeches that might contain multiple topics are only counted towards a single topic in the model. The result is that, in some cases, the "true" number of topics addressed by MEPs is under-represented and an inflated zero count is observed. In order to account for the inflated zero count, we model MEP speech-making as a two-stage process using a zero-inflated negative binomial regression model [6]. A zero-inflated negative binomial model includes a Logit regression component to capture the binary process determining whether or not a MEP speaks on a topic, and a negative-binomial regression component that seeks to capture the count process determining the number of speeches made, given that a MEP has chosen to speak on a topic.

In order to explain the variation observed in our dependent variable, we include variables relating to MEP's ideology, voting behavior, and the institutional structures in which they find themselves embedded within. We account for the left-right ideological position of a MEP's national party (as a proxy for MEP ideology) using data from [21]. Following [17], we also include a measure of how often MEPs vote against their party group in favor of their national party and vice versa. The idea behind including these variables is that MEPs rebelling against one party in favor of another will either try to explain such behavior in their speeches thus increasing the count, or hide their behavior by making no speeches, thus decreasing the count. These data were taken from an updated version of the [10] dataset provided by those authors. In order to capture an MEP's committee positions we include dummies for committee membership, chairs, and Rapporteurs in committees that are directly related to a given topic. Committees were manually matched with topics to achieve this. We control for whether or not an MEP serves in the Parliamentary leadership. Controls are also included for the total number of speeches made by an MEP and the percentage of MEP speeches that are available in English as these are liable to affect the observed MEP speech count. Finally, we also include dummy variables to control for an MEP's country of origin, EP party-group membership, and the topic on which they are speaking. All institutional and control variables were scraped from the legislative observatory of the European Parliament.

The regression presented in Fig. 8 provides further validation for

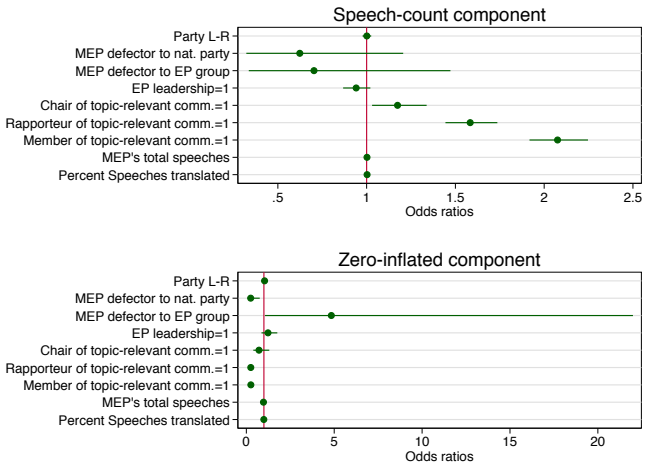


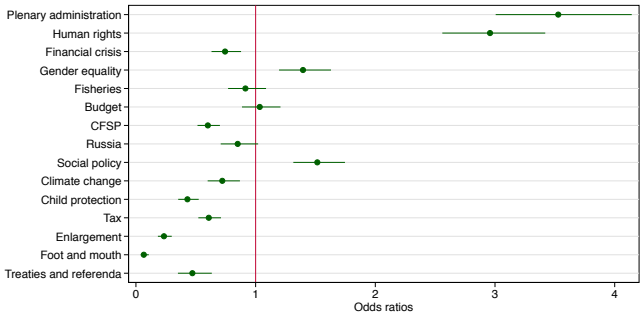
Figure 8: Plot of coefficients for regression model.

the results of our topic modeling approach. The coefficients of the model have been exponentiated so as to represent odds ratios and aid interpretation. For the Logit component of the model accounting for zero inflation, an exponentiated coefficient value above 1 implies that an increase in that covariate leads to an increase in the odds that a zero is observed (no speech is made), while any value below 1 implies an increase that variable leads to a decrease in the odds of a zero being observed (a speech being made). For the count component of the model exponentiated coefficient values above 1 are interpreted as implying a positive relationship between the predictor and outcome variable, while values below 1 imply a negative relationship between the predictor and outcome variable.

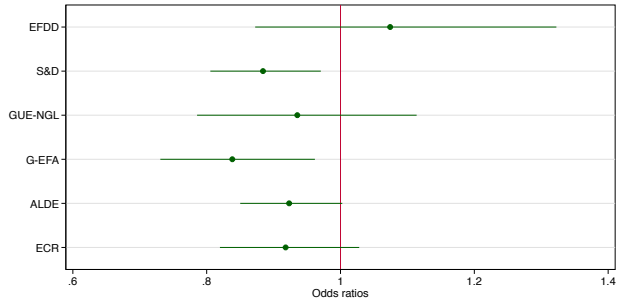
We begin with the zero-inflated component of the model in Fig. 8. The model suggests that a MEP's national party ideology impacts upon whether or not they make speeches on a given topic, with more right-wing MEPs tending to make no topic speeches more often than left-wing MEPs. Furthermore, MEPs defecting to national parties tend to make speeches more often than those not defecting, while the opposite is true for MEPs defecting to EP groups from national parties. This is in line with the findings of [17] who demonstrate that MEPs who rebel against their European party groups tend to make more speeches explaining why they do so, while those rebelling against their national party tend to make less speeches advertising their defection from the national party majority.

Of the institutionally related variables, holding a leadership position or a chair of a topic relevant committee has no significant relationship to MEP speech making, while being a member of a committee relevant to a topic, or holding a Rapporteurship for such a topic-relevant committee significantly impact upon whether or not MEPs make a speech that topic. The odds that a MEP makes no speeches on a given topic decrease by a factor of 0.255 if a MEP is a Rapporteur of a topic-relevant Committee and decrease by a factor of 0.259 if that MEP is a member of a topic-relevant committee. The results also show that the odds of an MEP making no speeches on a given topic decrease for MEPs that make more speeches in total. The result relating to the percentage of MEP speeches that are in English (whether translated or originally so) is also found to be significant, suggesting that MEPs with more speeches available in English tend to make speeches on a given topic more often.

Moving to the speech-count component of the model, the results further reinforce our expectations that MEP positions within the Parliamentary committee system impact upon how much attention they pay to a particular topic. When an MEP holds a committee



(a) Topic fixed effects topic



(b) Party group fixed effects

Figure 9: Fixed effects from regression model.

chair, Rapporteurship, or committee membership relevant to a particular topic, the odds that said MEP will make a speech on that topic increase by a factor of 1.173, 1.582, and 2.077 respectively.

In order to clarify the substantive size of the effects found in the model, Fig. 9(a) plots the odds ratio of different topics that entered into the regression model but were not displayed in Fig. 8. To plot these fixed effects odds rations, we treat the Euro crisis topic as the baseline. As can be seen, there are significant differences observed between the prevalence of different topics. The most prevalent topic is related to administrative matters in the plenary, and the odds of this topic appearing in an MEP speech are about 3.5 times greater than the odds of a speech relating to the Euro crisis topic. This is not surprising given that administrative matters frame all discussions in the plenary. Perhaps more surprising is the prevalence of the human rights topic relative to the other topics in the analysis. The odds of a speech relating to human rights is about 3 times greater than the odds that a speech relates to the Euro crisis. The relative prevalence of this topic suggests that MEPs regularly comment on human rights issues. Indeed, when one delves into the speeches appearing in this topic, a broad concern for violations of human rights across different contexts is evident. The relative prevalence of topics such as gender equality and social policy is also noteworthy, and suggests that the Parliament actively debates such issues despite the fact that the EU has little formal legislative competencies in these areas.

Fig. 9(b) plots the odds ratios associated with different party groups within the Parliament, treating the European People's Party (EPP) group as the baseline. As can be seen, there is some variation in the odds that a speech on a given topic emerges from a given party, but most of these differences are not statistically significant. This result reflects the fact that speech time is distributed between party groups according to their relative size. Both the *Progressive Alliance of Socialists and Democrats* (S & D) and the *European Greens–European Free Alliance* (G–EFA) groups are found to differ from the EPP group in terms of the odds a speech on a given topic comes from them. The odds of a topic speech being from either of these groups is less than the odds of a topic speech being from the EPP by a factor of just over 0.1.

6. CONCLUSIONS

In this paper, we have proposed a new two-layer matrix factorization methodology for identifying topics in large political speech corpora over time, designed to identify both niche topics related to events at a particular point in time and broad, long-running topics. We applied this method to a new corpus of all $\approx 210k$ English language plenary speeches from the European Parliament during a 15 year period. In terms of providing substantive insight into the political processes of the European Parliament, the topic modeling method has allowed us to unveil the political agenda in the Parliament, and the manner in which this has evolved over the time period under consideration. By considering three distinct case studies, we have demonstrated the distinctions that can be drawn between the day-to-day political work of the Parliament in policy areas such as fisheries on the one hand, and the manner in which exogenous events such as economic crises and failed treaty referenda can give rise to new topics of discussion between MEPs on the other. Once the Parliamentary agenda was extracted from the corpus of speeches, we explored the determinants of MEP attention to particular topics in the 7th sitting of the Parliament. We demonstrated how MEP ideology and voting behavior affect whether or not they choose to contribute to a topic, and once such a decision has been made, we demonstrated how the committee structure of the Parliament structures MEP contributions on a given topic.

The initial insights provide by the dynamic topic modeling approach presented in this paper demonstrate the potential of these methods to uncover the latent dynamics in MEP speech-making activities and thus allow for new insights into how the EU functions as a political system. Much remains to be explored in terms of the patterns in political attention that emerge from the topic modeling approach. For instance, one would expect that political attention might well translate into influence over policy outcomes decided upon in the Parliament. Tracing influence to date has been difficult, as a macro-level picture of where and on what topics MEP attention lies has been unavailable. Linking political attention to political outcomes would help to unveil who gets what and when in European politics, which is a central concern for a political system that is often criticized for lacking democratic legitimacy. this is but one direction in which future research might proceed.

While this paper has focused on European Parliament speeches, the proposed topic modeling method has a number of potential applications in the study of politics, including the analysis of speeches in other parliaments, political manifestos, and other more traditional forms of political texts. It is also generally appropriate in domains where large-scale, longitudinal text corpora are naturally represented in discrete segments.

Acknowledgments. This research was partly supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

7. REFERENCES

[1] G. Benedetto. Rapporteurs as legislative entrepreneurs: the dynamics of the codecision procedure in europe's parliament. *Journal of European Public Policy*, 12(1):67–88, 2005.

[2] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proc. 23rd International Conference on Machine Learning*, pages 113–120, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] C. Boutsidis and E. Gallopoulos. SVD based initialization: A head start for non-negative matrix factorization. *Pattern Recognition*, 2008.

[5] S. Bowler and D. M. Farrell. The organizing of the european parliament: Committees, specialization and co-ordination. *British Journal of Political Science*, 25(02):219–243, 1995.

[6] A. C. Cameron and P. K. Trivedi. *Regression analysis of count data*, volume 53. Cambridge university press, 2013.

[7] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. In *NIPS*, 2009.

[8] D. Greene, G. Cagney, N. Krogan, and P. Cunningham. Ensemble Non-negative Matrix Factorization Methods for Clustering Protein-Protein Interactions. *Bioinformatics*, 24(15):1722–1728, 2008.

[9] J. Grimmer. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35, 2010.

[10] S. Hix, A. Noury, and G. Roland. Dimensions of politics in the european parliament. *American Journal of Political Science*, 50(2):494–520, 2006.

[11] C. B. Jensen, S.-O. Proksch, and J. B. Slapin. Parliamentary Questions, Oversight, and National Opposition Status in the European Parliament. *Legislative Studies Quarterly*, 38(2):259–282, 2013.

[12] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–91, 1999.

[13] C. Lin. Projected gradient methods for non-negative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[15] D. O'Callaghan, D. Greene, J. Carthy, and P. Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications (ESWA)*, 2015.

[16] S.-O. Proksch and J. B. Slapin. Position taking in European Parliament speeches. *British Journal of Political Science*, 40(03):587–611, 2010.

[17] S.-O. Proksch and J. B. Slapin. *The politics of parliamentary debate: Parties, rebels and representation*. Cambridge University Press, 2014.

[18] K. M. Quinn, B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228, 2010.

[19] T. Raunio. Parliamentary questions in the European Parliament: Representation, information and control. *The Journal of Legislative Studies*, 2(4):356–382, 1996.

[20] A. Saha and V. Sindhwani. Learning evolving and emerging topics in social media: A dynamic NMF approach with temporal regularization. In *Proc. 5th ACM Int. Conf. Web search and data mining*, pages 693–702, 2012.

[21] R. Scully, S. Hix, and D. M. Farrell. National or European Parliamentarians? Evidence from a New Survey of the Members of the European Parliament. *JCMS: Journal of Common Market Studies*, 50(4):670–683, 2012.

[22] J. B. Slapin and S. O. Proksch. Look who's talking: Parliamentary debate in the European Union. *European Union Politics*, 11(3):333–357, 2010.

[23] M. Steyvers and T. Griffiths. Probabilistic Topic Models. In T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2006.

[24] R. Sulo, T. Berger-Wolf, and R. Grossman. Meaningful selection of temporal resolution for dynamic networks. In *Proc. 8th Workshop on Mining and Learning with Graphs*, pages 127–136. ACM, 2010.

[25] Q. Wang, Z. Cao, J. Xu, and H. Li. Group matrix factorization for scalable topic modeling. In *Proc. 35th SIGIR Conf. on Research and Development in Information Retrieval*, pages 375–384. ACM, 2012.

[26] N. Yordanova. The Rationale behind Committee Assignment in the European Parliament Distributive, Informational and Partisan Perspectives. *European Union Politics*, 10(2):253–280, 2009.