Combining clusters from network analysis with structural topic modelling: a late-fusion approach

Rafael Mesquita[1], Antonio Henrique Pires dos Santos

(Department of Political Science, Federal University of Pernambuco - Brazil)

NB: This working paper is ~2/3 complete. In full form, we would like to include a second test application of the method, perhaps using social media data (e.g., Twitter graph + content). We submit the paper to PNOC for both feedback on the manuscript and for partnership opportunities with other researchers who might be interested in joining with the data.

Please do not cite or circulate further without the authors' permission.

Abstract (200w)

This paper introduces a new method for combining clustering information from network analysis with textual information from Structural Topic Models. This approach can be classed under the broader family of "consensus" or "cluster ensembles" methods, that is, solutions that combining multiple partitions of a same dataset, arrived at by different methods, into a unified partition. We explore a late-fusion variant by which data are first classed independently according to network and textual information, producing two alternative partitions, which are then fused into one. After motivating the approach through a didactic example, we apply the solution to one test case: a corpus of all 17,913 resolutions adopted by the UN General Assembly since 1946. We first apply network community detection and identify 2,901 communities based on the pattern of cross citations. Then, we apply Structural Topic Model to the textual content of the corpus, outlining 40 topics. We then combine both sets to arrive at 125 definite clusters. The application reveals that the method can improve the quality of group partitioning by enabling the expansion of network communities to include singletons that share textual content with larger clusters.

Keywords: network analysis, community detection, structural topic model, text as data, cluster analysis, cluster ensemble, UN

1 Introduction

This paper introduces a method to combine clustering information from network analysis with that from Structural Topic Models (STM) to synthesize both inputs into a new global partition. The need for such an operation arises from the fact that entities can be observed and quantified with respect to multiple features, inherent as well as relational, and the affinities arising from each set of attributes might not be identical. Two election candidates might be very distinct from one another considering the content of their speeches, but equal in others such as the provenance of their campaign funds.[2] Similarly, any pair of international

[1] Corresponding author: rafael.mslima@ufpe.br

[2] Compare for instance Moreira (2020) and Mancuso et al. (2021).

organizations might be similar with regards to the provisions in their founding documents and regiments, and yet be disjoint when considering who are the states that place themselves under such obligations (Jetschke et al. 2021).

With the progress of computational social sciences and growing popularity of text-as-data, natural language processing, and network models among students of politics, the field seems ripe to engage with the problem of how to weave these separate threads of information into unified solutions. Network analysis offers a comprehensive canvas for this task. It allows us to consider agents not in isolation but with a view towards their relational dynamics (Knoke and Yang 2008). Further, networks can be enhanced by adding data on attributes inherent to the actors. The resulting "attributed graphs" are thus a rich data structure that offers abundant material to chose from so as to find patterns among entities.

The approach presented in this paper consists in first performing clustering procedures separately and then merging them into one. That is, for a same corpus, deriving a set of *topics* through STM, a set of *communities* via network analysis, and later combining them into a single *clustering* solution that takes into consideration information from both partitions. It is therefore an application of the so-called "late fusion" approaches to aggregating the results from distinct algorithms.

Such solutions are well-known in data science and in more applied quantitative domains of other disciplines, often classed as "cluster ensemble" or "consensus" classification problems (for a review, see Chunaev 2020; Boongoen and Iam-On 2018). They have been seldom used in Political Science and International Relations, however. We believe this paper can contribute not only by presenting its own application of this approach, but also as a general introduction to this class of aggregation procedures for the fields of Political Science and International Relations.

We begin by establishing the motivation for the method with a pedagogical example on the problem of classification in section 2. Section 3 presents the solution to the problem by explaining the late fusion approach. Then, we provide a real-world test case with a corpus of UN General Assembly resolutions. [to be completed]

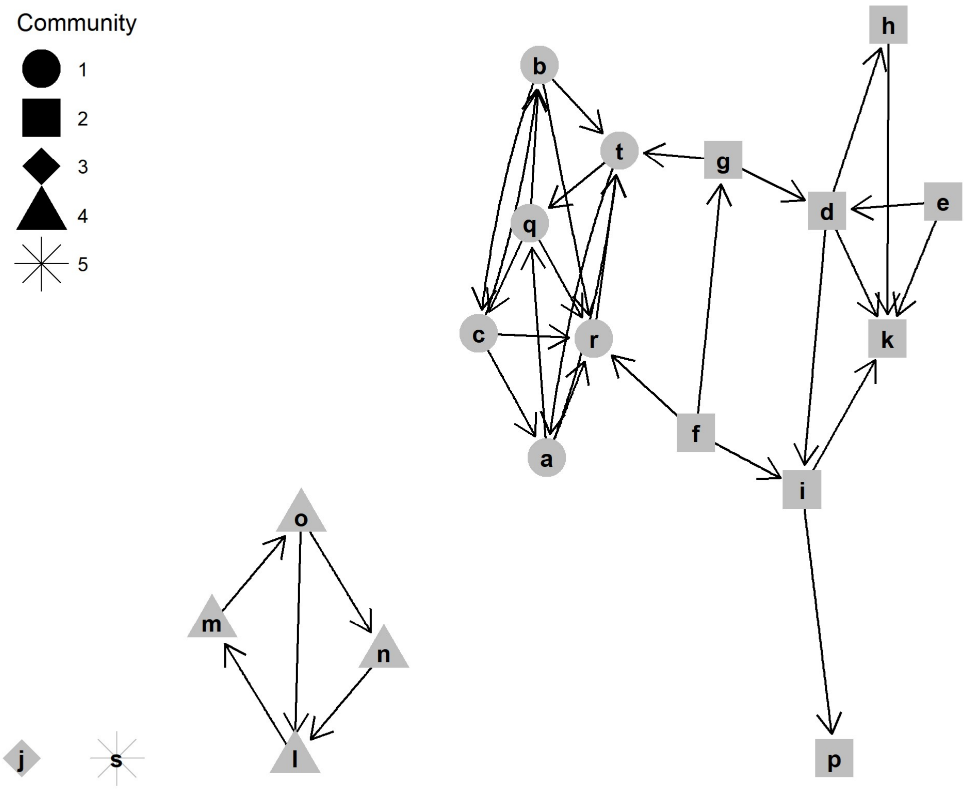
2 Motivation

Let a network or graph *G* be defined as the triplet *G = (N, M, A)*, wherein *N* correspond to the set of nodes, *M* to the set of edges connecting the nodes, and *A* to the attributes of the nodes. Nodes can be scored across multiple attributes. For our motivating example, let us consider a network *G1* of documents that are interconnected, for instance pieces of legislation that cite each other. *G1* has *N* = 20 texts and *M* = 33 directed edges between them.

The node and edge information can be used to divide the graph into communities. The goal of community detection is to partition a larger network into smaller and coherent subdivisions (Newman 2018). The procedure involves three important notions. First, there is more than one way to define what constitutes a community, so that available algorithms pursue different quality functions. Second, the node is the entity to be classified. This means that group membership becomes in a way a new attribute that is mapped on top of the nodes. Third, most available implementations produce non-overlapping subgroups. With few exceptions[3], the algorithms that are available in popular statistical software will place nodes into exclusive groups, so that it is not possible for a node to be simultaneously a member of two communities, nor to be "X% present" in a number of groups. Submitting *G1* to community detection produces a partition, that is, a solution which places each of the *N* nodes into one of *C* communities.

By applying the *Infomap* algorithm to the network in this example (Rosvall and Bergstrom 2008)[4], we arrive at the partition in Figure 1, which indicates each of the five resulting communities via node shapes.

Figure 1: Example network, with network communities only



Source: elaborated by the author.

Community #1 has six nodes with dense interconnection. Community #2 is larger, with eight members, but sparser links. Community #4 contains four nodes that only cite one

[3] Spectral clustering is a traditional solution that produces fuzzy membership (nodes are scored across a vector of underlying dimensions). Other approaches in stand-alone packages to produce shared membership for nodes include clique percolation (Lange 2021). For a survey of approaches, see Fortunato and Hric (2016).

[4] Infomap was chosen because, from the available algorithms in the *igraph* package for R, it is the only one that takes directionality into account.

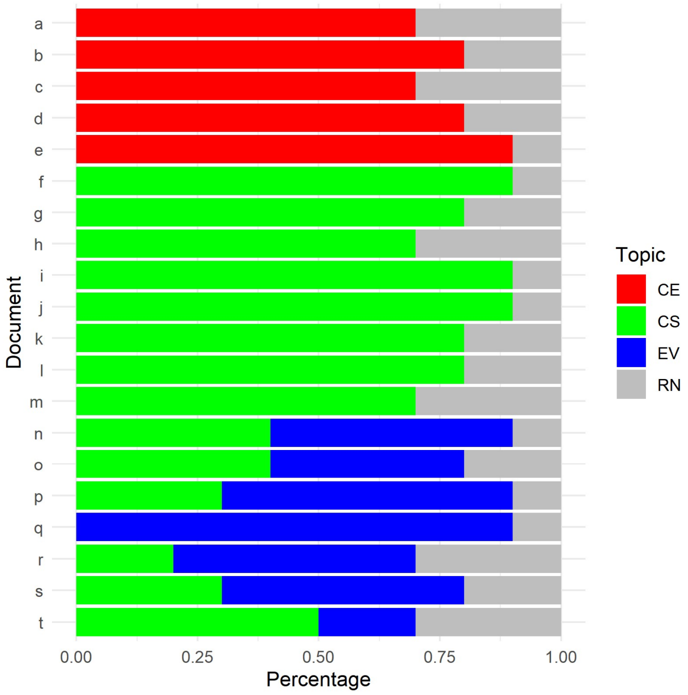
another and have no path connecting them to the larger component[5] formed by communities #1 and #2. Last, communities #3 and #5 are isolates, also called singletons, that neither cite nor are cited by any other document.

The proposed partition seems to have fulfilled to the dual task of approximating groups of nodes that are strongly connected to another and separating them from other nodes with which they have few or no links. It does not use any information on attributes to arrive at this partition.

We now factor in the textual content of the documents. Suppose that after submitting the 20 texts to STM we identify that their content can be classed into four topics. STM is an unsupervised text classification method that infers a set of topics from a corpus given its word usage and other covariates. It allows documents to be of mixed membership, that is, each document is regarded as a mixture from the set of proposed topics (Roberts et al. 2014; Grimmer, Roberts, and Stewart 2022). One of the output of the model is a matrix of document-level scores indicating what is the relative participation of the text's content within each of the four topics. That is, for each node *n* there is a vector of length equal to the number of topics (four, in our example) with continuous values corresponding to the node's participation in each topic. The sum of the vector is equal to one, meaning that 100% of the content of the document was distributed across the topics.

Figure 2 below shows the distribution of the topics across 20 documents as stacked bars.

Figure 2: Example distribution of 20 documents across 4 topics



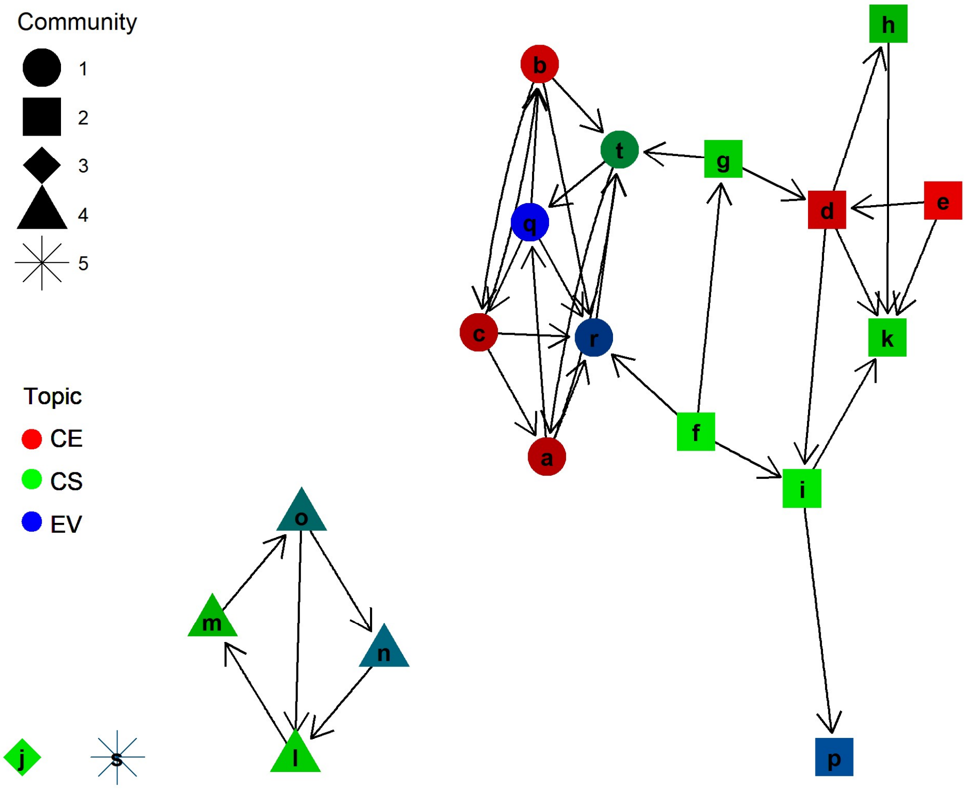
Source: elaborated by the authors.

[5] In network analysis, a component is defined as the set of nodes for which there is at least one path connecting all nodes within it. The network in Figure 1 has four components.

Indicated in grey, the topic “RN” stands for a random uniformly distributed topic. All documents touch upon it, in proportions between 0.1 and 0.3. Topic “CE” in red stands for a concentrated and exclusive subject. It only appears in documents *a* to *e* and in high concentration. Topic “CS” is also concentrated, as it occupies a large fraction of documents *f* to *m*, but it is shared across the majority of documents in the corpus. Last, topic “EV” is evenly shared across documents *n* to *t*, mostly with moderate values and seldom in a strong concentration. We could imagine topic RN as the formalities and conventions common to any document, no matter its subject matter, so that its information is not relevant to discriminate between texts. In contrast, CE is a strong discriminant: it is a rare subject, which only a few texts address with monothematic focus. CS, in contrast, stands as a more universal topic, mentioned by many texts at varying rates, and which appears in conjunction with other themes, such as EV.

The key feature of topic distribution is that documents are not exclusively assigned: texts have a weighted participation across all themes considered. The four topics can be assigned as attributes of the nodes (A_1 to A_4). Figure 3 updates our network graph, now overlaying each node with its proportional topic content. Node color refers to topic concentration, depending on how exclusively each document mentions EC (red), CS (green), or EV (blue).

Figure 3: Example network, with community and topic information



Source: elaborated by the authors.

It becomes clear that not all documents similar in content are similar in connections. Texts *j* and *s* are singletons (unconnected), but their texts are compatible with that of other, better-connected nodes. Likewise, some of the original groups were found to have greater content homogeneity (#4) while others cover all manner of topics (#2).

Typically, two documents on a same theme will cite one another, so that the groups resulting from text and network information should be similar. There are nonetheless reasons why this might fail to materialize and the groups will begin to differ: the author might be ignorant of past work and fail to make a reference, for instance.

If our goal is to arrive at a single partition that approximates documents based on their references as well as on the topics, we need a method to combine these strands of information.

3 Late fusion

In the survey conducted by Chunaev (2020), there are three broad options available to researchers that want to mesh attributes and network information. Early fusion methods combine the two data first to construct a modified network that is, in turn, submitted to classic community detection algorithms. Simultaneous fusion, for its part, injects attribute information into the community detection algorithm to customize its results. Late fusion approaches carry out community detection methods on network and on attribute data in parallel and only at the end choose a formula on how to combine both. As he and other authors have discussed, there are advantages and limitations to all these approaches and the choice hinges on the researcher's objectives and on the nature of the data.

Our focus on late fusion for the field of Political Science and International Relations is substantive and pragmatic. Substantively, the information conveyed, on the one hand, by textual content and, on the other hand, by relational ties, can at times vary. In some cases, the two sources might lead to similar results, so that their combination is merely a question of greater accuracy. In others, however, the circumstances of textual production can be very different from those leading to the formation of ties. Actors can rhetorically converge on a topic of widespread acclamation, but it does not follow that they will seek the same partners when pursuing it, as the example of UN discussions around democracy in the post-Cold War showed (Hecht 2017). We believe that this distance can also be of interest to researchers, so as to reveal more facets of the underlying dynamics studied. Late fusion allows to fully analyze each empirical thread and hence to make inferences on the congruence or discrepancy between both. Pragmatically, there is a greater abundance of statistical packages, as well as educational resources, on these two approaches as separate procedures. Because such implementations are readily available, late fusion is a less taxing way to combine them. While simultaneous fusion would require the user to make custom code (for instance to edit a quality function within network community detection algorithm so as to include note attributes), late fusion takes the output of "off the shelf" algorithms and combines them. We believe the latter situation is representative of the majority of users, who should be able to use statistical packages as they come and would like to add some complementation to their results. This also leads us to one of the strengths of our method: it is agnostic with regard to community detection algorithms as well as topic models. The method works regardless of whether communities were arrived at using, for instance, a label propagation or modularity maximization approach for community detection, or whether topics were found through STM, LDA, or BERTopics. The approach only requires a node partition for each, so that it can be employed to any algorithm.

In our approach, we consider the network communities to be the primary result and consult node attribute data as a complement to hone the original findings. If follows that this combination does not treat both steps of the process equally: the final partition is expected to have more in common with the network communities, amended to take text into consideration. The intuition is that the network community results should satisfactorily cover content distribution as well, and only incidentally leave out a relevant document that should be later brought into the original groups.

Late fusion applies to our motivating example as follows. Table 1(a) is a contingency table: rows indicate network communities, columns the four topics, and cells are a simple sum of topic content. Formally, for a community *C* and a topic *A*, the cell value for their crosstabulation will be:

$$x_{ca} = \sum_{1}^{n_c} t_a$$

Where t_a corresponds to the score on topic *A* for each one of the n_c nodes inside community *C*.

Hence, community #1 has five documents and, by adding the weight of each of these 5 documents on topic RN (0.3, 0.2, 0.3, 0.1, 0.3, 0.3), we arrive at 1.5. It is thus a measure of the volume of each topic per community. This tabulation provides a first clue on what a given community "is about", but it might be misleading since this measure is affected by community size. This is why in Table 1(b) we divide cell values by community size. Precisely:

$$x_{ca} = \frac{\sum_{1}^{n_c} t_a}{n_c}$$

The cells now indicate the thematic concentration within each community.

Table 1: Contingency table

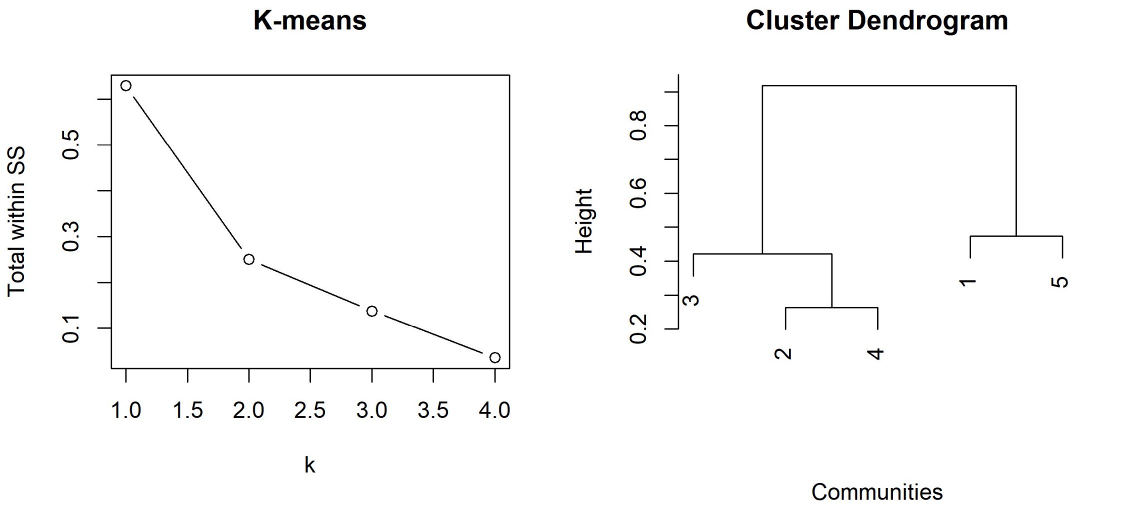
(a) **Community #** **(size)**	RN	CE	CS	EV	*Sum*		**(b)**	RN	CE	CS	EV	*Sum*
#1 (6 nodes)	1.5	2.2	0.7	1.6	*6*		**#1 (6)**	0.2	0.4	0.1	0.3	*1*
#2 (8)	1.3	1.7	4.4	0.6	*8*		**#2 (8)**	0.2	0.2	0.6	0.07	*1*
#3 (1)	0.1	0.0	0.9	0.0	*1*		**#3 (1)**	0.1	0.0	0.9	0.0	*1*
#4 (4)	0.8	0.0	2.3	0.9	*4*		**#4 (4)**	0.2	0.0	0.6	0.2	*1*
#5 (1)	0.2	0.0	0.3	0.5	*1*		**#5 (1)**	0.2	0.0	0.3	0.5	*1*
Sum	*3.9*	*3.9*	*8.9*	*3.6*	*20*		*Sum*	*0.9*	*0.6*	*2.5*	*1.1*	*5*

Source: elaborated by the authors.

Table 1(b) suggests that community #1 was devoted to CE, primarily, and to EV, secondarily. Communities #2 and #4, in turn, had a sizeable share of their content on CS. Given these thematic priorities, we have some guidance on what other communities we could fuse with #1, #2 and #4 to create more homogenous and comprehensive sets. Notably, we should like to bring communities #3 and #5, which are isolates, under communities #2 and #1, respectively, given their contents.

Our approach considers the network communities C_1 to C_5 as the new fundamental entities and their concentration of topics A_1 to A_4 as attributes or features. This transforms our question into a typical clustering problem, albeit starting from a higher-order, aggregate entity. The matrix represented by Table 1(b) can be submitted to classic clustering solutions that take rectangular matrices as inputs. Hierarchical clustering, for instance, can be used to reveal which network clusters are closer to one another regarding their thematic distributions, whereas k-means can suggest a new set of clusters that will minimize internal heterogeneity. These applications are exemplified in Figure 5.

Figure 5: K-means and dendrogram for example data



Source: elaborated by the authors.

Although both options could be explored, we will continue with k-means for conciseness. By considering each of the five communities as an entity and the four topics as variables that describe them, k-means will attempt to group the communities into a smaller number of sets (C^k<C) that minimizes the sum of squared distances.

It is clear to see that this operation is a form of division of the larger partition (C) by the smaller (A), so that we can diminish the total number of network communities via successive incorporations that minimize group heterogeneity in the A attributes. The upper bound for the number of clusters K is given by C (communities stay as they are) and the lower bound is 1 (all communities are fused into one). We will refer to the result of this division as *clusters* (K) to avoid confusion with the original communities (C) from network analysis.

If we choose K=3, the next step is to produce a membership table that associates the new clusters proposed by k-means to the 5 original communities from *Informap* and, consequently, places each of the 20 nodes into new clusters. The result of this updating is shown in Figure 6.

Figure 6: Final solution for example network, with clusters K=3

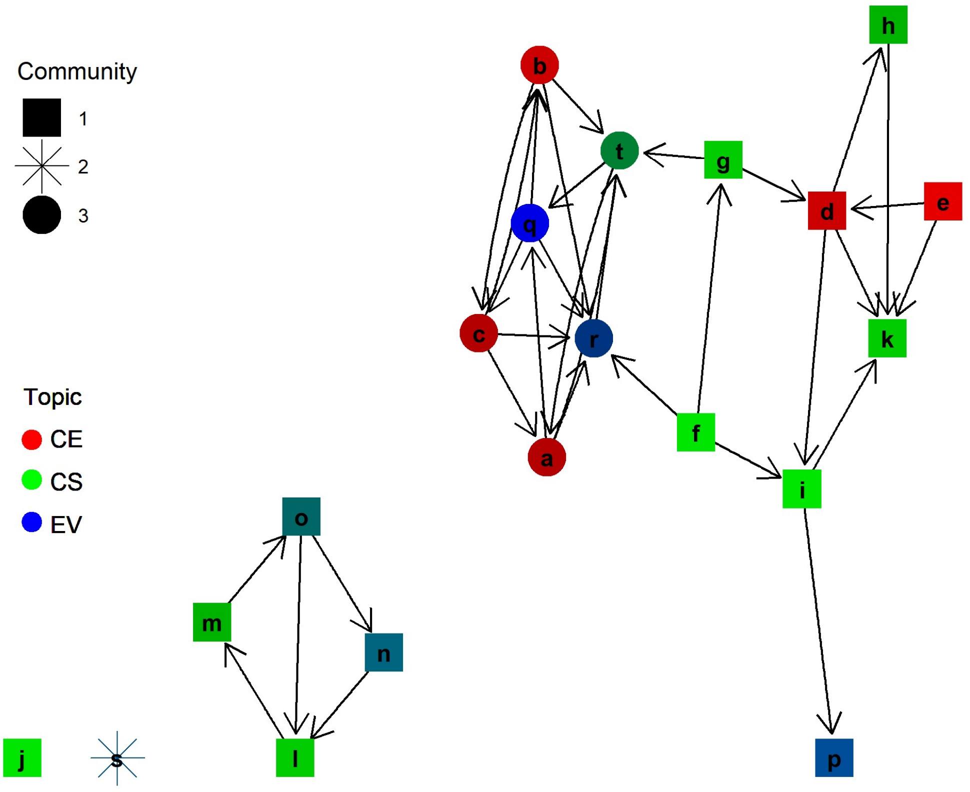
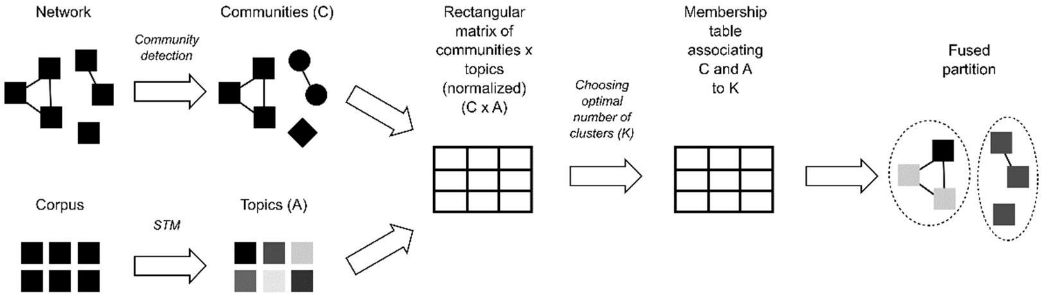


Figure: elaborated by the authors.

The new plot reveals that former communities #2, #3, and #4 were merged into a new cluster #1. Our goal was met, since many of the documents addressing CS were now brought under a single group. Former community #1 was not readjusted, but had we opted for K=2 instead of 3, the next step would have been incorporating document *s* into this community, creating a cluster with most of the EV topic.

The method can thus be summarized as in Figure 7.

Figure 7: Flowchart of the late fusion method


Network
Community detection
Communities (C)
Rectangular matrix of communities x topics (normalized) (C x A)
Choosing optimal number of clusters (K)
Membership table associating C and A to K
Fused partition
Corpus
STM
Topics (A)

Source: elaborated by the authors.

4 Test case: UN General Assembly resolutions

The *UN General Assembly Resolutions Texts (1946-2019)* dataset by Mesquita and Pires (unpublished) is a citation network as well as a corpus. It contains the integral text of all resolutions adopted by the UNGA from 1946 to 2019, along with relational information concerning citations between documents. Since these resolutions can be interpreted as a synthesis of majority sentiment in world politics across multiple issues, cluster detection is a way to inductively outline the main themes of global governance in the post-Second World War (for a review, see Mesquita and Pires 2022). As per our method, we first detect network communities and STM topics separately in order to, at a final stage, fuse both partitions.

4.1 Network communities

From the 17,913 resolutions, 2,206 are unconnected (i.e., neither citing nor receiving citations). This means that the UNGA citation network is similar to other networks characterized by a giant connected component that hosts the majority of entities (14,591 resolutions, or 81% of the whole corpus) surrounded by a corona of isolates (resolutions that have no connections) and tendrils (chains of a few resolutions that do not connect to the largest component).

For such networks, the typical choice is to discard the isolates and apply community detection only to the giant component. This is because community detection algorithms will place each isolate in their own community, thus producing a long-tailed distribution of many communities of size 1. This is the ideal setting for our approach, since we can expect many such documents to address a shared theme but, for an exogenous reasons, fail to form appropriate connections. Consider, for instance, resolutions for accepting credentials of officials and national representatives. Such resolutions do not need to cite other references, since they refer to an immediate administrative decision, but employ almost unvarying language to authorize this action. The lexical similarity would allow us to locate these “distant cousins”, pull them out of their isolation, and group them under a new cluster.

We apply *igraph’s* walktrap algorithm to the network with walk length of 20 steps, which was the length that gave us the highest modularity (0.851). We had a total of 2,901 communities, of which 695 were inside the giant component. Most importantly, community size followed a long-tailed distribution, in which 76% of the communities had only 1 resolution, as shown in Table 2 below:

Table 2: Size of the communities in the UNGA citation network

Community size	(0-1] resolutions	(1-10]	(10-100]	(100-1,000]	(1,000-2,886]
Number of clusters	2,206	529	152	10	4

Source: elaborated by the authors.

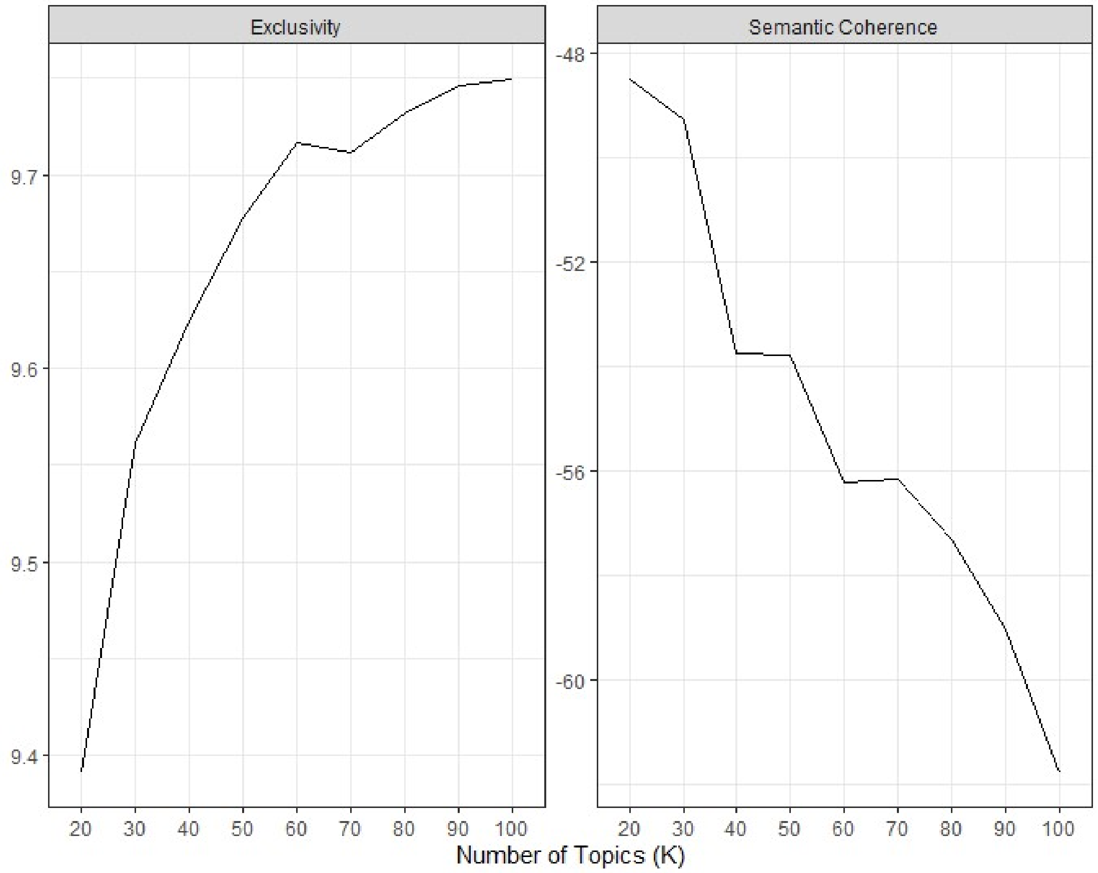
Starting from the upper bound of C=2,901, we must scale down the number of communities, trying to bring isolates to participate in larger groups. For that, we use the textual information of the resolutions.

4.2 Structural Topic Models

In this section, we present the details of how we performed topic modeling. Before applying STM, the resolutions were tokenized and the following transformations were made in the corpus: removal of punctuation, numbers, symbols and URLs, conversion to lower case, stemming, removal of stop words, words with less than two characters, and words appearing in less than 100 documents.

To run the STM, it is necessary to assign a number of topics manually, but there are some strategies to choose optimally. First, in the *stm* R package, when setting the initialization type to "Spectral" and specifying K = 0, STM selects a number of topics based on the algorithm of Lee and Mimno (2014). In our case, the number chosen by the algorithm was 75. This number cannot be considered the definitive one, but it offers a place to start. Secondly, we may use the criteria of semantic coherence (Mimno et al. 2011) and exclusivity (Bischof and Airoldi 2012) to select the best model. We ran the function search to retrieve these measures for models with number of topics ranging from 10 to 100 with intervals of 10. Results for exclusivity and coherence for these runs are shown in Figure 8.

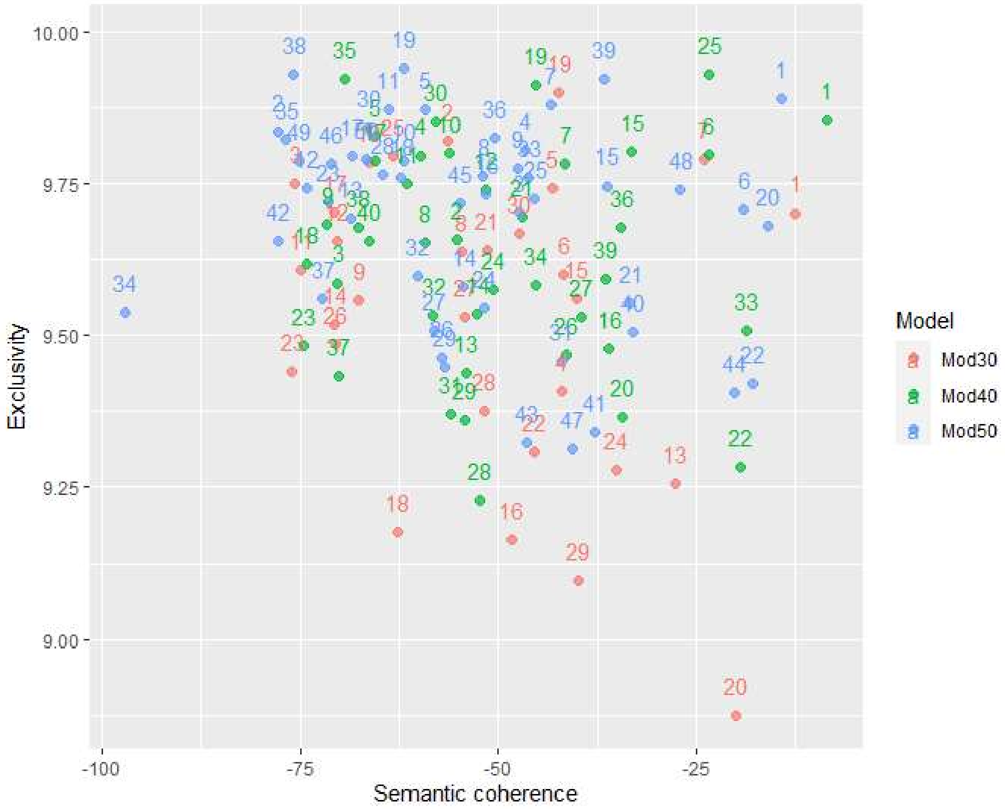
Figure 8: Diagnostic values by number of topics (K = 20 to 100)



Source: Elaborated by the authors.

After analyzing these outcomes, we ran four models with numbers of topics equal to 30, 40, 50 to compare the quality of topics again based on those measures, displayed in Figure 9. Finally, we performed a qualitative inspection of the topics' top words to assess the gains and losses of each model in terms of content. After these steps, we chose to run the model with 40 topics.

Figure 9: Quality of topics in models (K = 30 to 50)



Source: Elaborated by the authors.

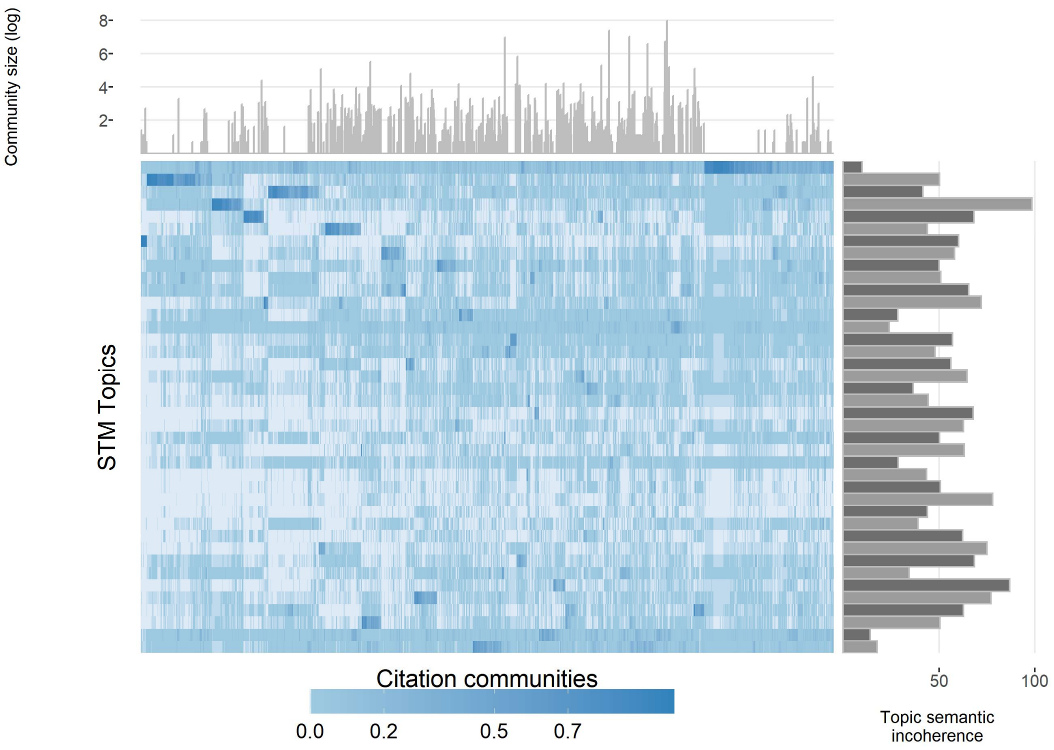
4.3 Late fusion

The first step in our late fusion approach is to assemble a contingency table such as Table 1 that will allow us to see STM topic occurrence per network community. Since the combination of topics and clusters leads to a 40 x 2,901 table, we cannot use an actual table as before, so that we will represent it via a heatmap instead.

The heatmap in Figure 10 shows communities as columns and textual topics as rows. As there are many topics and communities, we omit labels at first. Cell color represents the ratio we proposed in Table 1(b), that is, the share of each topic normalized by community size. Hence, the darker the shade, the more monothematic a community.

The columns and rows are sorted according to their Euclidean distances, which is a straightforward way to identify patterns via hierarchical clustering. We add two adjacent plots to the heatmap. The plot on the top refers to community size, in log scale, i.e., how many resolutions are inside the community identified by each column. The plot on the right refers to the semantic coherence of each textual topic. Semantic coherence is a quality metric for topics from STM that is high when the words expected under a topic co-occur frequently in the same documents and low when they do not co-occur often.

Figure 10: Heatmap of topic concentration per community

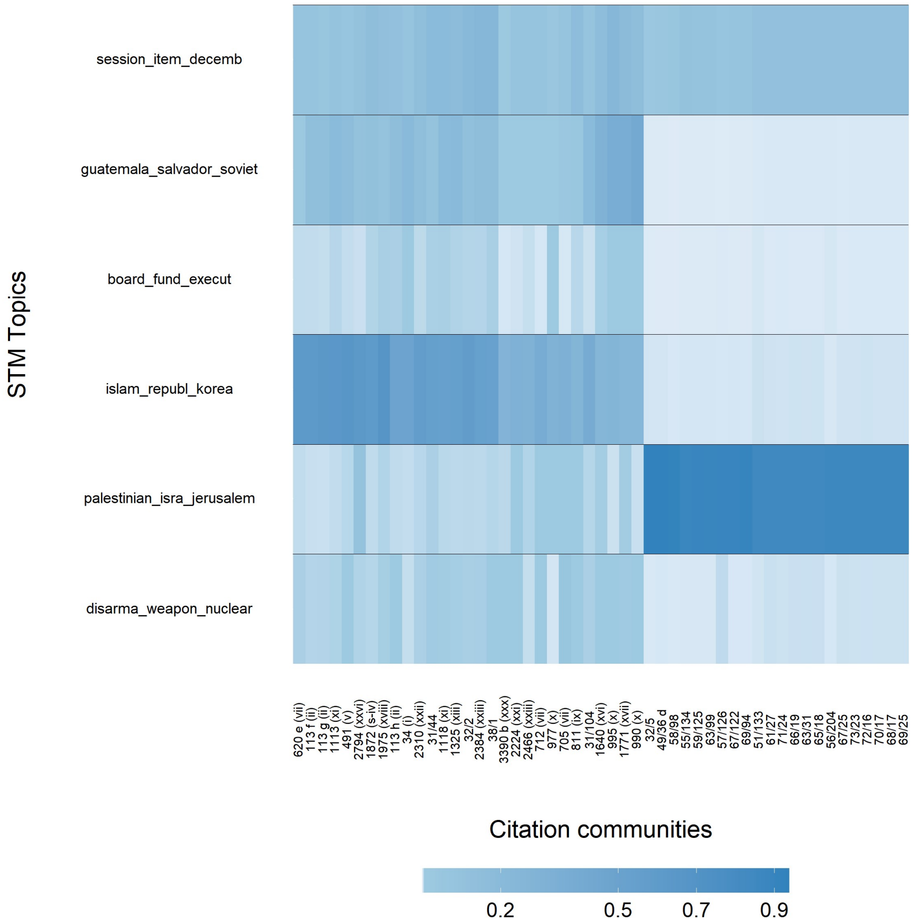


Source: elaborated by the authors. Note that semantic coherence scores are originally on a negative scale. We therefore took the liberty of reporting absolute values on the plot and, for consistency, name the axis by the opposite notion of "incoherence". Hence, a topic with a low bar is coherent and one with a high bar is incoherent.

The row and column order proposed by hierarchical clustering is suggestive of some patterns. The upper part of the heatmap presents, on the right and left corners, darker stripes, which should be interpreted as communities that are concentrated in single topics. By crossing this information with the bar plot on the top panel, we infer that these concentrated communities are also small. By looking at the bar plot on the right panel, we further conclude that some are coherent, such as the one on the top row. These are potentially singletons or, at most, tendrils, that address a same topic and could potentially be combined in larger clusters.

A qualitative appraisal is necessary to judge if the topics spanning these eligible communities are meaningful. Figure 11 zooms in on the upper left portion of the previous plot and adds labels for both axes. The row labels correspond to the top 3 most frequent words in each STM topic. For the column labels, we selected the most cited resolutions inside each citation cluster as a quick heuristic to indicate us the subject that centers each community.

Figure 11: Inset from Figure 10



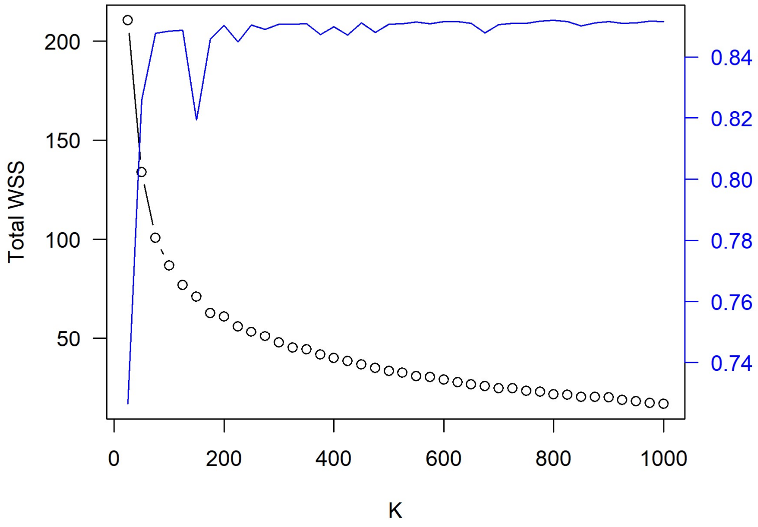
Source: elaborated by the authors.

The segment of citation communities starting from the community centered on resolution 620 E (VII) and ending on the community around 990 (X) is strongly concentrated on the themes summarized by keywords/stems “Guatemala”, “Salvador, “Soviet” and “Islam”, “Republ”, “Korea”. Looking into the content of these communities, we see that resolutions such as “Admission of New Members” (620 E [VII]), “Question of the Representation of China” (990 [X]), and others in this segment deal with admission of member states. This matches with the composition of the topics, which are made up mostly of proper nouns corresponding to country names. As we move to the left of the plot, we see that the community centered on resolution 32/5 marks a thematic change. This community of 16 resolutions contains drafts on Israel and Palestine, which matches with the textual topic keywords. Other clusters in this segment, such as the ones centered on resolutions 58/98, 57/126, 51/133 are singletons. These highly concentrated singletons are good candidates for merger into a larger unit.

To choose the number of clusters, we can again compare the results for different values of K in k-means clustering or the dendrogram. We do not reproduce the dendrogram because, with 2,901 leaves, it is not visually helpful, but its inspection has indicated that a cut at height 0.2 would lead to 969 clusters, at 0.3 to 448, and at 0.4 to 213. If we compare a similar range of k-means clusters, starting at 1000 and ending at 25, decreasing 25 steps at a time, we obtain

Figure 12 below. The left vertical axis indicates total within sum of squares, as a measure of quality for the k-means clusters, and the right vertical axis indicates network modularity[6] if the proposed partition were overlayed to the original nodes, as a measure of quality for network communities.

Figure 12: Within sum of squares and modularity for different number of clusters K



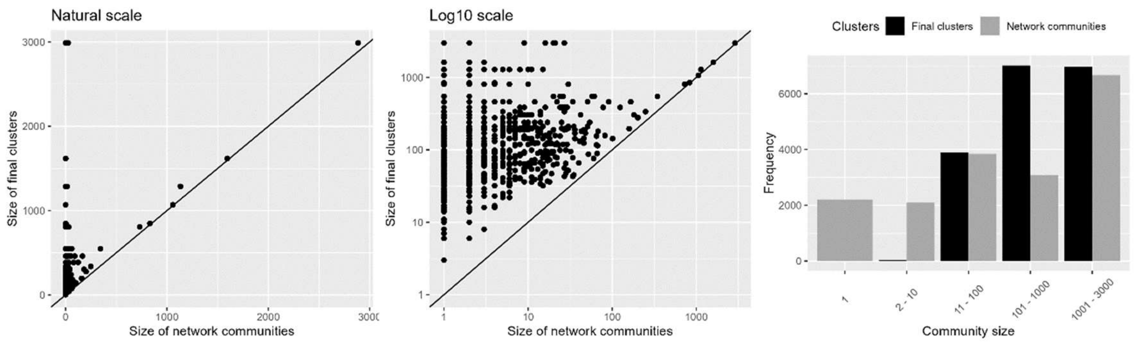
Source: elaborated by the authors.

The so-called elbow method would recommend choosing K either at the fifth or seventh points of the series (K=125 or 175). Either would conserve a modularity score (0.849, 0.846) similar to that obtained by the original community detection that yielded 2,901 communities (0.851). We opt therefore for 125 clusters and regroup the original nodes under the new partition.

After overwriting the previous communities affiliations with the final clusters derived from k-means, the resolutions from the original network are now regrouped into a new partition. We can compare community size before and after the late fusion so as to verify if our goal of regrouping singletons was achieved. Figure 13 displays three comparisons.

Figure 13: Comparison of community and cluster sizes

[6] "Modularity is a measure of the extent to which like is connected to like in a network. It is strictly less than 1 and takes positive values if there are more edges between nodes of the same type than we would expect by random chance. It can also take negative values if there are fewer such edges than we would expect by chance" (Newman 2018, 205).



Source: elaborated by the authors.

The left pane is a scatter plot wherein each point is a resolution, the x-coordinate corresponds to the size of the community to which it belonged before the fusion, and the y-coordinate represents the size of the new community in which the resolution was placed. Hence, if a resolution sits on top of the diagonal line, its new community was just as big as the old one, but if it is placed high above the line, this means that this resolution moved from a small community to a larger one. The plot reveals that many of the resolutions that originally were in small communities were moved into larger clusters. The dots on the left corner indicate that some resolutions that before were in communities approximately of size 1 were placed in the largest existing clusters, of sizes close to 3,000 and 1,500. The center pane shows the same data on a logged scale for better visualization. It demonstrates that, apart from the incorporation of isolates into big clusters, resolutions previously on communities of slightly greater sizes, between 10 and 100, were also regrouped into bigger clusters. Last, the right pane is a histogram that compares how many resolutions were placed in clusters of size 1, 2-10, 11-100, 101-1,000, and 1,001-3,000 in both partitions. It shows that in the network communities, approximately 2,000 resolutions were in singleton communities, and that the same amount was in communities of size 2 to 10. In the final clustering solution, there were no singletons, very few documents in communities of size less than 10, and the bulk of resolutions was sorted into clusters of size ranging from 101 to 1,000.

[to do: second test case, conclusion]

References

Bischof, Jonathan M., and Edoardo M. Airoldi. 2012. "Summarizing Topical Content with Word Frequency and Exclusivity." In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 9–16. ICML'12. Madison, WI, USA: Omnipress.

Boongoen, Tossapon, and Natthakan Iam-On. 2018. "Cluster Ensembles: A Survey of Approaches with Recent Extensions and Applications." *Computer Science Review* 28 (May): 1–25. https://doi.org/10.1016/j.cosrev.2018.01.003.

Chunaev, Petr. 2020. "Community Detection in Node-Attributed Social Networks: A Survey." *Computer Science Review* 37 (August): 100286. https://doi.org/10.1016/j.cosrev.2020.100286.

Fortunato, Santo, and Darko Hric. 2016. "Community Detection in Networks: A User Guide." *Physics Reports* 659 (November): 1–44. https://doi.org/10.1016/j.physrep.2016.09.002.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press.

Hecht, Catherine. 2017. "Advantages and Disadvantages of Inclusive Multilateral Venues: The Rise and Fall of the United Nations General Assembly Resolution on New or Restored Democracies." *International Politics* 54 (6): 714–28. https://doi.org/10.1057/s41311-017-0058-4.

Jetschke, Anja, Sören Münch, Adriana Rocío Cardozo-Silva, and Patrick Theiner. 2021. "Patterns of (Dis)Similarity in the Design of Regional Organizations: The Regional Organizations Similarity Index (ROSI)." *International Studies Perspectives* 22 (2): 181–200. https://doi.org/10.1093/isp/ekaa006.

Knoke, David, and Song Yang. 2008. *Social Network Analysis*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc. https://doi.org/10.4135/9781412985864.

Lange, Jens. 2021. "CliquePercolation: An R Package for Conducting and Visualizing Results of the Clique Percolation Network Community Detection Algorithm." *Journal of Open Source Software* 6 (62): 3210. https://doi.org/10.21105/joss.03210.

Mancuso, Wagner Pralon, Rodrigo Rossi Horochovski, Ivan Jairo Junckes, and Neilor Fermino Camargo. 2021. "Pragmatismo ou ideologia?setores empresariais e financiamento de campanha em 2014." *E-legis* 14 (34): 29–49.

Mesquita, Rafael, and Antonio Pires. 2022. "What Are UN General Assembly Resolutions for? Four Views on Parliamentary Diplomacy." *International Studies Review* 25 (1): viac058. https://doi.org/10.1093/isr/viac058.

Mimno, David, and Moontae Lee. 2014. "Low-Dimensional Embeddings for Interpretable Anchor-Based Topic Inference." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1319–28. Doha, Qatar: Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1138.

Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. "Optimizing Semantic Coherence in Topic Models." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–72. EMNLP '11. USA: Association for Computational Linguistics.

Moreira, Davi. 2020. “Com a Palavra Os Nobres Deputados: Ênfase Temática Dos Discursos Dos Parlamentares Brasileiros.” *Dados* 63 (1): e20180176. https://doi.org/10.1590/001152582020204.

Newman, M. E. J. 2018. *Networks*. Second edition. Oxford, United Kingdom ; New York, NY, United States of America: Oxford University Press.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. “Structural Topic Models for Open-Ended Survey Responses.” *American Journal of Political Science* 58 (4): 1064–82. https://doi.org/10.1111/ajps.12103.

Rosvall, Martin, and Carl T. Bergstrom. 2008. “Maps of Random Walks on Complex Networks Reveal Community Structure.” *Proceedings of the National Academy of Sciences* 105 (4): 1118–23. https://doi.org/10.1073/pnas.0706851105.