Simonetti *et al. Applied Network Science* (2025) 10:5
https://doi.org/10.1007/s41109-025-00693-z

Applied Network Science

RESEARCH Open Access

Statistically validated network for analysing textual data

Andrea Simonetti[1*], Alessandro Albano[1,5], Michele Tumminello[1] and T. Di Matteo[2,3,4]

*Correspondence:
andrea.simonetti01@unipa.it

[1] Department of Economics, Business, and Statistics, University of Palermo, Viale delle Scienze, Ed. 13, 90128 Palermo, Italy
[2] Department of Mathematics, King's College London, The Strand, London WC2R 2LS, UK
[3] Complexity Science Hub Vienna, Josefstädter Straße 39, 1080 Vienna, Austria
[4] Centro Ricerche Enrico Fermi, Via Panisperna 89A, 00184 Rome, Italy
[5] Sustainable Mobility Center (Centro Nazionale per la Mobilità Sostenibile-CNMS), Milan, Italy

Abstract

This paper presents a novel methodology, called Word Co-occurrence SVN topic model (WCSVNtm), for document clustering and topic modeling in textual datasets. This method represents the corpus as a bipartite network of words and documents to rigorously assess the statistical significance of word co-occurrences within documents and document overlap based on shared vocabulary. By employing the Leiden community detection algorithm to the SVN, distinct communities of words can be identified and interpreted as topics. Similarly, documents can be sorted into groups based on their thematic similarities. We demonstrate the effectiveness of our approach by analyzing three datasets: a set of 120 Wikipedia articles, the arXiv10 dataset, which consists of 100,000 abstracts from scientific papers, and a sampled subset of 10,000 documents from the original arXiv10. To benchmark our results, we compare our approach with several well-established models in the field of topic modeling and document clustering, including the hierarchical Stochastic Block Model (hSBM), BERTopic, and Latent Dirichlet Allocation (LDA). The results show that WCSVNtm achieves competitive performance across all datasets, automatically selecting the number of topics and document clusters, whereas state-of-the-art methods require prior knowledge or additional tuning for optimization. Finally, any advancements in community detection algorithms could further improve our method.

Keywords: Document clustering, Topic modeling, WCSVNtm, Bipartite networks, Statistically validated networks (SVN), BERTopic, hSBM, LDA

Introduction

In the realm of text mining, two significant scientific challenges revolve around document clustering and topic modeling, both aimed at extracting valuable insights from large collections of text. Topic modeling techniques were developed to automatically uncover latent themes in the documents, enabling more efficient information retrieval, content summarization, and trend analysis. Document clustering, on the other hand, focuses on grouping similar documents together based on their content, facilitating the organization, exploration, and analysis of textual data.

This paper addresses both challenges by proposing a novel approach, WCSVNtm, which integrates the network representation of textual data with the Statistically Validated Networks (SVN) approach (Tumminello et al. 2011). This integration enables the

simultaneous grouping of similar documents (document clustering) and the extraction of latent themes (topic modeling).

In the context of existing approaches, Latent Dirichlet Allocation (LDA) (Blei et al. 2003) has become the state-of-the-art for topic modeling due to its flexibility and effectiveness. However, as highlighted by Gerlach et al. (2018), LDA lacks a systematic method for determining the number of topics, leading to uncertainty in model selection (Belford et al. 2018). Indeed, without empirical justification, the use of the Dirichlet prior limits the model's adaptability to real-world text characteristics such as Zipf's law (Zipf 1936; Piantadosi 2014). Unfortunately, the LDA extensions do not fully solve the mentioned issues due to Dirichlet priors, model structures, or heuristic parameter optimisation limitations.

Neural network approaches for conducting topic modeling have recently been introduced (Dieng et al. 2020). Word embedding (Bengio et al. 2003; Mikolov et al. 2013a, b) revolutionised natural language processing by capturing the semantic relationships between words in a continuous vector space. In topic modeling, word embedding offers a modern approach to inferring topics by clustering words with similar embedding. By leveraging the semantic coherence included in word embeddings, the construction of topic representations can effectively capture the underlying thematic structures within the corpus. Building on this foundation, document embeddings have emerged as an extension of word embeddings in topic modeling tasks. Modern approaches such as BERTopic leverage the power of document embeddings to identify topics by clustering semantically similar documents. While word embeddings capture the semantic relationships between individual terms, document embeddings provide holistic representations of entire texts by aggregating word-level information into a unified vector. Then, by applying clustering algorithms to document embedding spaces, it becomes possible to group semantically similar documents, facilitating the effective organisation and exploration of textual data (Angelov 2020). Moreover, community detection methods from network science have also been used for the clustering of document embeddings (Altuncu et al. 2019, 2021). However, the effectiveness of these models heavily relies on the quality of the document embeddings generated. Embedding models trained on non-specialized datasets may struggle to represent domain-specific language or rare terms, leading to suboptimal clustering in specialized corpora and impacting the clustering results (Gururangan et al. 2020; Xu and Lapata 2019).

Knowledge Graph is another method to manage textual data, making use of ontology as a schema layer (Paulheim 2017), representing word entities and their relationships. These models demonstrate noteworthy applications in fields such as genomics and systems biology (Gramatica et al. 2014). However, knowledge graph methods are compelling in some applications, they require a predefined ontology schema for practical training. These constraints present challenges when applying these models to domain-specific or small datasets.

In complex networks, a shift from heuristic to probabilistic models, similar to topic modeling, is seen in community detection (Fortunato 2010). While the two approaches are usually used to address different problems (Airoldi et al. 2008), both fields share conceptual similarities (Airoldi et al. 2008; Ball et al. 2011; Lancichinetti et al. 2015). Community detection aims to identify groups of nodes with similar connectivity patterns,

revealing network structure and potential functional units (Fortunato 2010). Other methods have been proposed with the goal of finding community structures in bipartite networks, prompting a move towards probabilistic inference approaches such as stochastic block models (SBMs) (Airoldi et al. 2008; Holland et al. 1983; Karrer and Newman 2011), mirroring the evolution in topic modeling.

Gerlach et al. (2018) introduced a unified approach to topic modeling and community detection. They turn the problem of inferring topics into a problem of inferring communities by representing the word-document matrix as a bipartite network. The authors showed that clustering and finding topics in collections of written documents can be tackled using hierarchical Stochastic Block Models (hSBM). The results of their work showed that hSBMs outperform and overcome many of the difficulties of the LDA. Similarly, Hyland et al. (2021) investigated the task of clustering and finding topics from a collection of documents for which additional information (metadata and hyperlinks between documents) is available. The authors extended the work of Gerlach et al. (2018) by showing how the metadata and hyperlinks can be incorporated in the same framework by using multilayer hSBMs (Zhu et al. 2013; Valles-Catala et al. 2016; Peixoto 2015). Specifically, they add, in the bipartite network of document-word, a hyperlink layer connecting the different written documents and a metadata-document layer that incorporates tags.

In this paper, for the first time, the network representation of textual data is integrated with the Statistically Validated Networks (SVN) approach (Tumminello et al. 2011) to face document clustering and topic modeling. Specifically, we present a new method, called WCSVNtm, which represents texts in a bipartite network of words and documents, and then employs Statistically Validated Networks, to statistically test the significance of each link in a projected weighted network. We show that this approach effectively filters out irrelevant connections among words through an exact statistical test. Subsequently, we apply a community detection algorithm to identify both word communities (topics) and clusters of similar documents.

In the framework of topic models, Simonetti et al. (2023) demonstrated that SVN can be effectively utilised to measure topic coherence, showing a significant correlation with human judgment. However, their study employed an LDA model to retrieve topics. In contrast, our paper proposes to identify topics using an SVN-based method directly. This approach allows us to uncover topics with high coherence naturally. Moreover, the modularity contribution of each community (topic) can be interpreted as a measure of coherence since it is an intensive quantity that assesses the tendency of words within a given topic to occur in the same sentences jointly. With this approach, we study words' semantic similarities, discovering the latent topics behind a collection of texts.

The proposed method follows the same path of Gerlach et al. (2018) and Hyland et al. (2021) by representing textual data as a bipartite network but differs in the network construction, which incorporates a battery of statistical tests with appropriate corrections for multiple comparisons, such as Bonferroni (Miller 1981) and False Discovery Rate (FDR) (Benjamini and Hochberg 1995) corrections. We first capture the significant word co-occurrences, considering the heterogeneity of words. Then, once we find the main connections among words, we introduce a threshold to retrieve the strongest similarities among documents represented through these meaningful word co-occurrences.

Finally, the topic-document association is carried out through the Fisher's exact test (Sprent 2011). We evaluated our method's ability to perform both document clustering and topic modeling. Specifically, we evaluated the method on three datasets: a set of 120 Wikipedia articles, the arXiv10 (Farhangi et al. 2022) dataset, which consists of 100 000 abstracts from scientific papers, and a sampled subset of 10 000 documents from the original arXiv10. For both document clustering and topic modeling, we compared our method to hSBM and BERTopic, and for topic modeling specifically, we also compared it to LDA. The results demonstrate that our method shows competitive performance with state-of-art methods when applied to datasets of size spanning four orders of magnitude, highlighting its robustness and scalability.

The structure of the paper unfolds as follows. First, we present the proposed methodology in detail. Next, we demonstrate the practical application of our method to real-world data, comparing its performance against the state-of-the-art models. Finally, we engage in a thorough discussion of the findings, followed by concluding remarks.

Methodology

The present section illustrates how the proposed method, WCSVNtm, can be used to face the tasks of document clustering and topic modeling. Our proposal consists of the following four steps (visually represented in Fig. 1):

1. SVN projection on the set of words
2. SVN projection on the set of documents
3. Document clustering and topic modeling
4. Documents - Topics association

The initial step, involving SVN projection on the set of words, follows the methodology outlined in Simonetti et al. (2023). However, all subsequent steps presented in this paper are introduced for the first time. The following subsections describe in detail the steps of the proposed methodology.

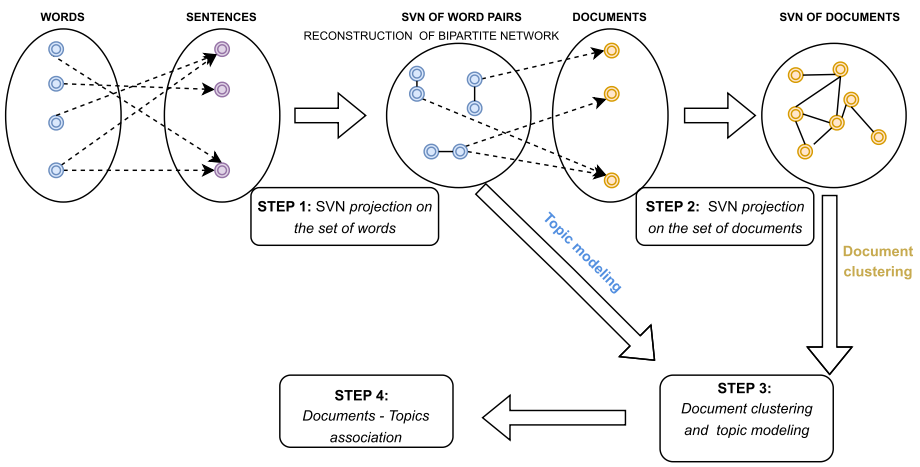


Fig. 1 Illustration of the WCSVNtm method

SVN projection on the set of words

In the first step, we use a modified version of the bag-of-words representation of a collection of documents, in which we split each document into sentences. Rather than employing the conventional document-term matrix representation, we opt for a sentence-term matrix representation. In this representation, we annotate a cell with 1 if a word occurs within a sentence. The idea is to construct a bipartite network from the sentence-term matrix, in which words and sentences are the two sets of nodes. Figure 2 shows an example of bipartite network, in which the set of nodes $S = \{s_1, s_2, s_3, s_4, s_5\}$ is made by five sentences, and the other set of nodes $W = \{w_1, w_2, w_3, w_4, w_5\}$ is made by the words that occur at least in two sentences.

A link is set between a word and a sentence if that word belongs to that sentence. Once we construct the bipartite network, we use the Statistically Validated Network method (Tumminello et al. 2011) to obtain a projection on the set of words. The resulting projected network is a word-co-occurrence network (Zuo et al. 2016; Paranyushkin 2011).

The SVN method uses a statistical test to project only the links among words whose co-occurrences among sentences are statistically significant. To take into account the heterogeneity of the set of sentences, a suitable null hypothesis is considered, i.e., the number of sentences in which any two words co-occur is random. Indeed, a model of the null hypothesis is attained by performing a random rewiring of the network, in which the degree (heterogeneity) of words and sentences is kept constant as in the original data. The hypothesis test is constructed as follows: considering a corpus consisting of N sentences and the total number of times two words, w_i and w_j, occur individually in the sentences of the corpus as N_i and N_j, respectively (see Fig. 3). Then, let us indicate with X_{ij} the random variable that counts the times the two words co-occur jointly in the same sentences. We want to statistically validate the co-occurrences of the words w_i and w_j against a null hypothesis of random co-occurrence that accounts for the heterogeneity of the considered words. The probability distribution under the null hypothesis that

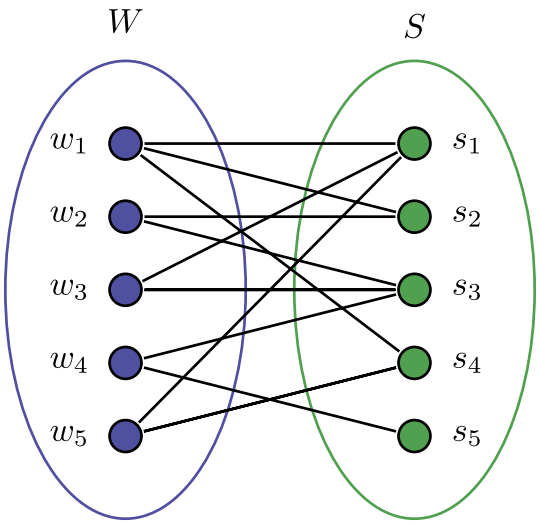


Fig. 2 Example of bipartite network where S is a set of 5 sentences and W is the set of 5 words. A link between a sentence and a word indicates that the word belongs to the sentence

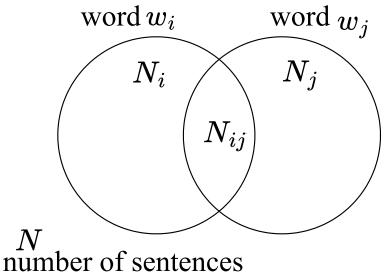


Fig. 3 Venn diagram showing the overlap of number of sentences, N_{ij}, where words w_i and w_j occur. N_i and N_j indicate the number of sentences in which word w_i and w_j occur, respectively

describes the random co-occurrence is the Hypergeometric distribution, according to which the probability of observing $X_{ij} = N_{ij}$ co-occurrences is given by

$$\mathrm{pmf}_H(X_{ij} = N_{ij}|N, N_i, N_j) = \frac{\binom{N_i}{N_{ij}}\binom{N - N_i}{N_j - N_{ij}}}{\binom{N}{N_j}}, \tag{1}$$

where parameters N_i and N_j naturally allow for incorporating the aforementioned heterogeneity of words in the null hypothesis.

The distribution introduced is used to test if an excess of the number of co-occurrences between any pair of words, w_i and w_j, occurs. Indeed, assuming that the number of observed co-occurrences of these words is N_{ij}, the probability that a value larger than or equal to N_{ij} is observed by chance, according to the null hypothesis, is:

$$p_v(N_{ij}|N_i, N_j, N) = \sum_{X=N_{ij}}^{min(N_i,N_j)} \frac{\binom{N_i}{X}\binom{N - N_i}{N_j - X}}{\binom{N}{N_j}}. \tag{2}$$

To claim that the number of co-occurrences, N_{ij}, between two words is too large to be consistent with the null hypothesis of random co-occurrences, we must introduce a threshold α of statistical significance. However, since we are facing multiple and dependent comparisons, errors of the first kind are real issues. Therefore, we use the conservative Bonferroni correction (Miller 1981) for multiple hypothesis testing. The correction states that given a univariate threshold of statistical significance, α, then the threshold corrected for multiple hypothesis testing is $\alpha_T = \frac{\alpha}{T}$, where T is the total number of performed tests, be they dependent or otherwise. The advantage of the Bonferroni correction is that it provides a very strict control of the Family Wise Error Rate even when tests are dependent, as in this case, since the same word appears in many tests.

SVN projection on the set of documents

In the second step, we reconstruct a new bipartite network where the two sets of nodes are the documents and the word-pairs (links) in the statistically validated network, constructed previously. We set a link between a word-pair and a document if the pair occur in at least one sentence of the document. Figure 4 represents an

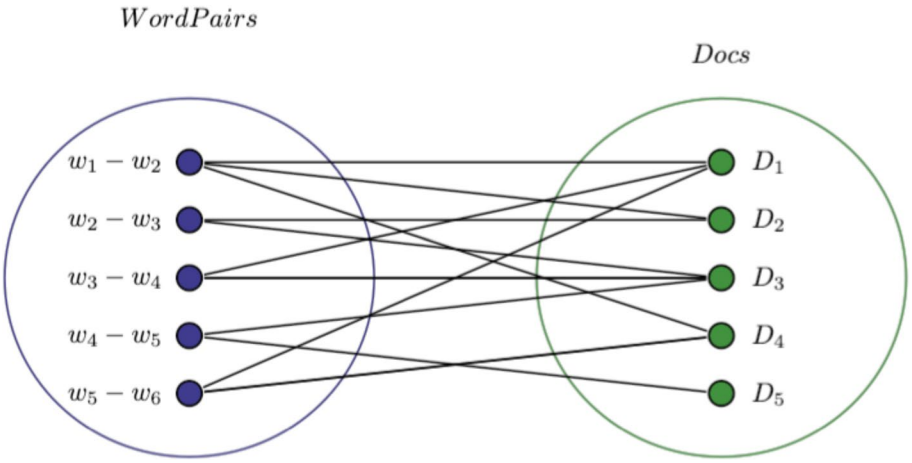


Fig. 4 Bipartite network of documents and statistically validated word-pairs

example of the new bipartite network of documents and word-pairs, in which the set of nodes $D = \{D_1, D_2, D_3, D_4, D_5\}$ is made by five documents, and the other set of nodes Word Pairs $= \{w_1 - w_2, w_2 - w_3, w_3 - w_4, w_4 - w_5, w_5 - w_6\}$ is made by five word-pairs that are previously statistically validated. Again, we apply the SVN method, but now we project on the set of documents and construct a validated network of documents. Since we face multiple and dependent comparisons, as in the previous step, we use again the Bonferroni correction (Miller 1981).

Document clustering and topic modeling

To perform both topic modeling and document clustering, we use the community detection algorithm described in Traag et al. (2019), called Leiden algorithm. Our method identifies communities[1] of documents by applying the Leiden algorithm to the statistically validated *network of documents*, while topics are derived by applying the same community detection algorithm to the *network of words*. In the document network, communities represent groups of similar documents, whereas in the word network, communities correspond to sets of semantically related words, which identify topics.

Additionally, we consider the link weights by incorporating the Pearson correlation coefficient as a word similarity metric, introduced for the SVN in Tumminello et al. (2013). The formula to compute the words correlation is

$$\rho(w_i, w_j) = \frac{N_{ij} - \frac{N_i N_j}{N}}{\sqrt{N_i(1 - \frac{N_i}{N})N_j(1 - \frac{N_j}{N})}}, \tag{3}$$

where N_i and N_j are the marginals, respectively of documents (words) w_i and w_j; N_{ij} is the number of common word-pairs (sentences) of i and j, and N is the total number of word-pairs (sentences).

[1] We use the term "community" to refer to groups of nodes that are detected as cohesive substructures within the network. Once these communities are extracted, they are treated as "clusters" of objects (e.g., words or documents), forming explicit partitions regardless of the network structure from which they originate.

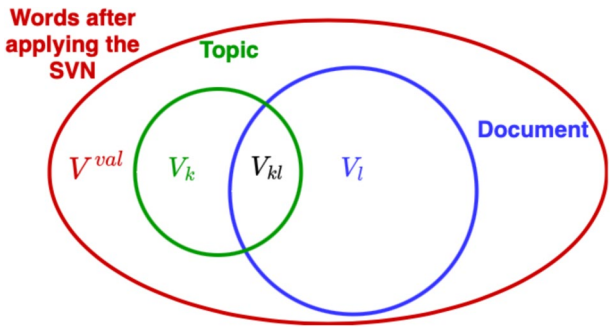


Fig. 5 Venn diagram describing the overlapping number of words, represented by V_{kl}, between a topic k and a document l, and their marginals V_k and V_l, respectively. The test for detecting over-expressions of topics in documents considers, as the universe, the words in the network obtained by applying the SVN, V^{val}

It is worth noticing that the Pearson correlation, ρ, is proportional to the Z-score of a Hypergeometric distribution (the test statistic value under a Z test (Casella and Berger 2002)) under the null hypothesis, where the constant of proportionality is $N^{\frac{1}{2}}$ (the total number of word-pairs) (Hatzopoulos et al. 2015).

Upon identifying a topic, we assign an importance score to each word within the topic to identify the most representative ones. For each word w_i, the score is calculated by dividing its modularity contribution $M(w_i)$ by the sum of the modularity contributions of all words within the community (topic). Specifically, the importance score $S(w_i)$ is calculated as

$$S(w_i) = \frac{M(w_i)}{\sum_{w_j \in C} M(w_j)},$$

where $M(w_i)$, the modularity contribution of word w_i, is given by:

$$M(w_i) = \frac{1}{2E} \sum_{j \in C} \left(A_{ij} - \frac{k_i k_j}{2E} \right),$$

where A_{ij} is the element of the adjacency matrix in the projected network of words, k_i and k_j are the degrees of nodes w_i and w_j, E is the total number of edges in the network, C is the community (topic) to which the word w_i belongs, and the sum is taken over all words w_j belonging to community C. This ensures that the importance score reflects the relative contribution of each word to the modularity of the topic.

Documents: topics association

After obtaining the topics, as communities in the word co-occurrence network, we evaluate their association with each document through the Fisher's exact test. This involves testing for over-expression using the Hypergeometric distribution. From Fig. 5, we identify an over-expression of a topic k in a document l if the number of overlapped words, V_{kl}, is statistically significant. To conduct this test, we count the shared words, V_{kl}, between a topic k and a document l, as well as their respective marginal counts, V_k and V_l, considering only the words within the resulting statistically validated network as the universe, V^{val}. As in the previous steps, the probability that a value larger than or equal to V_{kl} is observed by chance, according to the null hypothesis, is:

$$p_v(V_{kl}|V_k, V_l, V^{val}) = \sum_{X=V_{kl}}^{min(V_k,V_l)} \frac{\binom{V_k}{X}\binom{V^{val}-V_k}{V_l-X}}{\binom{N}{V_l}}. \quad (4)$$

To correct the multiple hypothesis testing, we use the False Discovery Rate (FDR) (Benjamini and Hochberg 1995), a statistical method helps control the expected proportion of false positives among all the rejected hypotheses. To implement the FDR correction, we apply the Benjamini-Hochberg procedure, which sorts the p-values in ascending order and compares each p-value to its corresponding threshold (based on the rank and the total number of tests). The FRD is a less stringent correction than the Bonferroni correction, used to control Type I errors (false positives), but is more effective in controlling Type II errors (false negatives). Although the FDR criterion is valid when the tests are independent, as Benjamini and Yekutieli (2001) pointed out, the controlling procedure also holds when the test statistics are positively dependent, which is the case under our analysis. We opt for the FDR criterion to promote greater associations between topics and documents. This method for document-topic association is an adaptation of the approach used to characterise communities in a network based on the attributes of their nodes, as presented in Tumminello et al. (2011) and Tumminello et al. (2023). In our case, the network communities correspond to topics (communities of words), while the attributes represent the documents in which each word appears.

Results

The methodology outlined in the previous section is employed on three datasets to perform document clustering and topic modeling tasks.

Firstly, we employed the 120 Wikipedia articles dataset (provided by Hyland et al. (2021)[2]). Each article is tagged with one of the following scientific categories: *Sub-fields of physics* (28 documents), *Branches of biology* (35 documents), and *Fields of mathematics* (57 documents), which together constitute the true partition of documents. Representing the dataset as a network, with documents as nodes and hyperlinks as directed edges, yields a network of 120 nodes and 309 edges (hyperlinks). This dataset is particularly well-suited for our purposes, as each document is labelled, facilitating direct comparison with the document clustering results in Hyland et al. (2021).

In our pre-processing procedure, we remove stop words since our method does not automatically discard them. Additionally, we opt for stemming over lemmatisation to reduce the various inflectional forms of words. The original dataset is already preprocessed; punctuation and capital letters have been removed. Therefore, to construct the bipartite network of words and sentences in Step 1, we generate synthetic sentences using non-overlapping sliding windows of fixed length to cover every single document. In addition, we employed two versions of the arXiv10 Benchmark dataset. This dataset covers a wide range of topics within scientific research, including 10 distinct categories, such as Computer Science, Physics, and Mathematics. The first version consists of a reduced subset of the dataset, containing 10 000 articles (1000 per class). The second

[2] https://github.com/martingerlach/hSBM_Topicmodel/tree/master/data

version includes the full dataset with 100 000 articles (10 000 per class). The aim is assessing the performance and scalability of our approach on a much larger and more complex set of documents.

For both document clustering and topic modeling, we compared our method to hSBM and BERTopic, and for topic modeling specifically, we also compared it to LDA. Evaluation on document clustering is carried out using metrics such as Maximum Partition Overlap (MPO) (Peixoto 2021) and Normalised Mutual Information (NMI) (Fano 1961). Details on these two metrics are provided in the Appendix. Broadly, NMI measures the global alignment of partitions, accounting for both matches and mismatches across them. MPO, on the other hand, focuses on the largest overlaps between partitions and may understate their quality if smaller overlaps are also significant. In both cases, a high value (close to 1) indicates strong agreement between the compared partitions.

As regards topic modeling, we evaluate the performance of several topic models on three different metrics: Coh_{SVN} (Simonetti et al. 2023), CV, and PMI (pointwise mutual information) (Röder et al. 2015). Finally, we examine document-topic associations, emphasizing their relationship with the document community structure, and providing an analysis of the primary topics. All analyses are performed in `Python`, through the packages `NLTK`, `NetworkX`, `Gensim` and `BERTopic` Grootendorst (2022), while the codes for hSBM provided by Hyland et al. (2021) are available at https://topsbm.github.io.

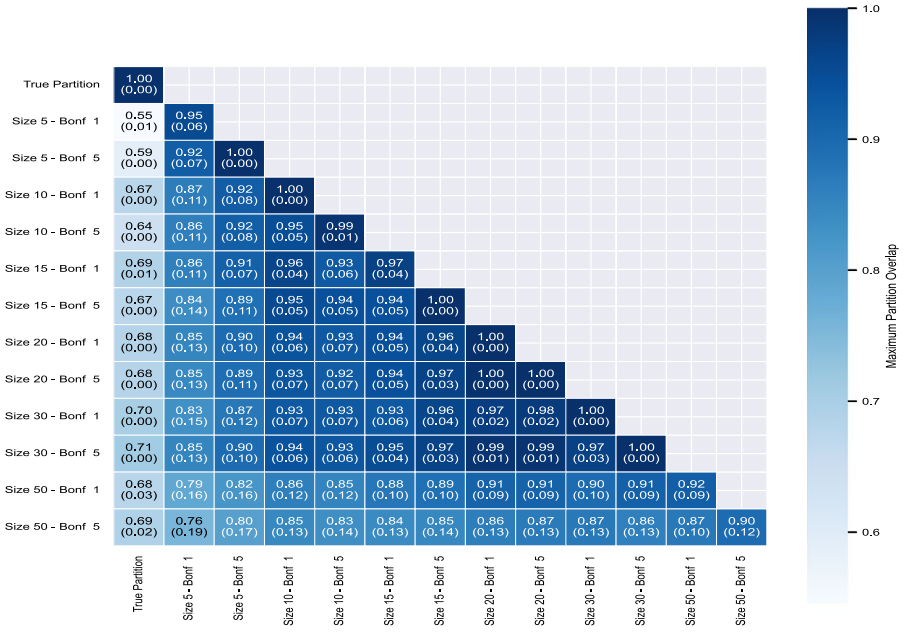
Selection of the best document community structure

Firstly, we perform the first two steps of our procedure, namely, SVN projection on the set of words (Step 1) and SVN projection on the set of documents (Step 2).

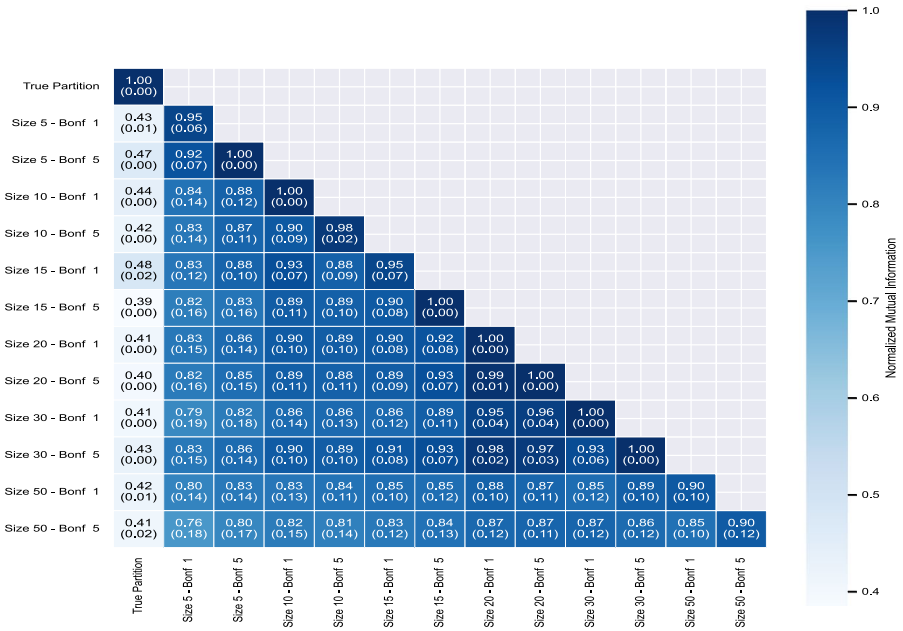
In the case of the Wikipedia dataset, we need to generate synthetic sentences using non-overlapping sliding windows of fixed length, as described in Step 1 of our methodology, to capture word co-occurrences within each document. This is necessary because the Wikipedia dataset had already been preprocessed and cleaned by Hyland et al. (2021). Therefore, to ensure a fair comparison, we followed the same approach and used the same processed data, which lack of punctuation. In contrast, for the arXiv10 Benchmark dataset, we directly use the existing sentences in the abstracts and titles of the articles. In this way these sentences serve as natural windows for detecting word co-occurrences. Nevertheless, to explore the sensitivity of the model to hyperparameters on a large-scale dataset and to demonstrate its robustness, we performed a sensitivity analysis also on this dataset (reported in the Appendix 3).

In Step 1, for the Wikipedia dataset, we consider different values of sentence length (5, 10, 15, 20, 30 and 50 words) and different values of the univariate threshold of statistical significance before the Bonferroni correction is applied, namely, $\alpha = 0.01, 0.05$. In Step 2, instead, we just set the value $\alpha = 0.01$ to be more conservative in validating the connections among documents. To determine the optimal values of the parameters (i.e., α and sentence length), we evaluate the retrieved partitions against the true partition for each combination of parameters. Given that the Leiden method is a heuristic algorithm, we replicate the experiment 100 times for each combination.

The heatmaps in Fig. 6 illustrate the similarity between different partitions as obtained at varying values of sentence length and threshold α. Figure 6a and b show the similarity



(a) *Maximum partition overlap measure.*



(b) *Normalized Mutual Information (NMI) measure.*

Fig. 6 Heatmaps. The average and standard deviation (in parenthesis) among 100 replicates using the MPO (**a**) and NMI (**b**) between and within different parameter settings. The scores represent the average over 100 replicates of applying the Leiden community detection algorithm on the SVN of documents to extract the document partition

scores obtained by using the MPO and the NMI, respectively. In both cases, the similarity values range between 0 and 1, where 1 indicates identical partitions and 0 represents completely dissimilar partitions. Darker shades of blue correspond to higher values of similarity, while lighter shades indicate lower similarity. The first column shows the overlap between the true partition and the partitions obtained in each configuration, providing a direct comparison of model outcomes with the ground truth. The main diagonal in the heatmaps does not exclusively consist of 1 s, since entries are calculated as averages over all pairs of the 100 independent realizations of the model. Among the different combinations, we select sentence length as 30 and $\alpha = 0.05$. This choice yields the highest score for the MPO measure (0.71) and a high score for the NMI measure (0.43). Additionally, this setting exhibits the lowest standard deviation when compared to the true partition for both measures. Notably, almost all combinations yield similar partitions, except for the sentence length of 5 words, underscoring the robustness of our proposed method in parameter selection.

Document clustering

In this section, we address the document clustering task, comparing our proposed method with the hSBM model and BERTopic. For the hSBM model, we evaluate its performance on both the bipartite network of documents and words, as well as the multilayer network that includes the hyperlink layer. Given that hSBM inherently produces a hierarchical clustering of documents, we analyse the results at two key levels: the highest non-trivial level (i.e., the first partition consisting of more than one cluster) and the second highest non-trivial level of the hierarchy.

Table 1 shows the similarity between the true partition and the outcomes of i) our model (using two levels of α: 0.01 in WCSVNtm_1 and 0.05 in WCSVNtm_5), ii) hSBM model, and iii) BERTopic. The similarity is evaluated in terms of both the MPO and NMI scores, and averages are calculated over 100 replicates of the system.

The second column in the table specifies the layers included in each model, such as document, word, or hyperlink layers, highlighting the underlying differences in the data utilised by each approach.

For the *Wikipedia 120* dataset, our method (WCSVNtm) demonstrates strong clustering performance. The model achieves a MPO of 0.70 (WCSVNtm_1) and 0.71 (WCSVNtm_5) with relatively low standard deviations (less than 0.01). In comparison, the hSBM model achieves a similar MPO (0.70) when incorporating the hyperlink layer. However, this result is somewhat misleading, as the hyperlink-based partition merges all documents into only two broad groups, failing to distinguish between Maths and Physics. When the hSBM model is applied to the document-word bipartite network, its performance drops significantly (MPO of 0.59 and NMI of 0.35), demonstrating its limitations in effectively leveraging text features alone. Including the hyperlink layer to the hSBM with Docs, Words layer does not improve the performance significantly. BERTopic, achieves the highest MPO score (0.72) and a competitive NMI score of 0.46. However, similar to the hyperlink-based hSBM, BERTopic identifies only two clusters, which significantly limits its ability to capture the true diversity of the dataset.

For the *arXiv 10k* dataset, WCSVNtm continues to exhibit robust performance, with WCSVNtm_1 and WCSVNtm_5 achieving MPO scores of 0.51 and 0.52, respectively,

Table 1 MPO and NMI scores between the models and the true partition. The values correspond to the mean over 100 replicates, with standard deviation in parenthesis

Models	Layers	MPO (Peixoto 2021)	NMI (Fano 1961)
Wikipedia 120			
WCSVNtm_1	Docs, WordPairs	0.70 (< 0.01)	0.41 (< 0.01)
WCSVNtm_5	Docs, WordPairs	0.71 (0.01)	0.43 (< 0.01)
hSBM	Hyperlinks	0.70 (< 0.01)	0.40 (< 0.01)
hSBM	Docs, Words	0.59 (0.10)	0.35 (0.12)
hSBM	Hyperlinks, Docs, Words	0.58 (0.09)	0.37 (0.03)
BERTopic	-	0.72 (< 0.01)	0.46 (< 0.01)
arXiv 10k			
WCSVNtm_1	Docs, WordPairs	0.51 (0.01)	0.47 (0.02)
WCSVNtm_5	Docs, WordPairs	0.52 (0.01)	0.47 (0.01)
hSBM	Docs, Words	0.43 (0.03)	0.39 (0.04)
BERTopic_refit	–	0.41 (< 0.01)	0.38 (< 0.01)
arXiv 100k			
WCSVNtm_1	Docs, WordPairs	0.41 (0.01)	0.35 (0.01)
WCSVNtm_5	Docs, WordPairs	0.42 (0.02)	0.38 (0.01)
hSBM	Docs, Words	0.44 (0.06)	0.37 (0.03)
BERTopic_refit	–	0.43 (< 0.01)	0.49 (< 0.01)

and consistent NMI scores of 0.47. The low variability in these scores highlights the stability of the method, even as the dataset size increases. In contrast, the hSBM model performs worse, with a MPO of 0.43 and an NMI of 0.39. On the other hand, BERTopic estimates a large number of clusters (over 400) for this dataset, leading us to speculate about possible ways to avoid over-fragmentation. For the sake of the present analysis, the easiest solution to this problem of BERTopic appeared to inform the model with the number of topics provided by our method. Then, to obtain the partition, for each document we select the topic assigned with the highest probability. We name "BERTopic_refit" this (refined) version of BERTopic. The results show that BERTopic_refit performance (MPO of 0.41 and NMI of 0.38) is comparable to hSBM, but still lower than the one of WCSVNtm.

Finally, for the largest dataset, *arXiv 100k*, our method continues to perform competitively, with WCSVNtm_1 achieving a MPO of 0.41 and an NMI of 0.35 and WCSVNtm_5 with MPO 0.42 and NMI 0.38. Interestingly, the hSBM model achieves higher scores for this dataset (MPO of 0.44 and NMI of 0.37). However, it should be noted that, to derive a partition of the corpus from the hierarchical tree generated by hSBM, we selected the third hierarchical level from the root. This specific choice ensures the maximum overlap with the original partition and thus represents an optimal selection for the hSBM. Importantly, this approach provides an advantage to hSBM over other method, as the threshold in the hierarchy was determined using prior knowledge of the "true" partition, a condition that would not be feasible in real-world scenarios where such knowledge is unavailable. Despite this advantage, the performance of our method remains within one standard deviation of that achieved by hSBM. BERTopic faces again the challenge of over-fragmentation, resulting in a large number of clusters. After

refining the model to avoid excessive fragmentation, BERTopic_refit achieves an MPO of 0.43 (comparable to other methods) and an NMI of 0.49. The NMI score is higher than those of both WCSVNtm and hSBM, showing its ability to effectively capture groups of similar documents in the dataset. However, it should be emphasized that the fragmentation issue remains a concern, as the model originally estimated more than 9 700 clusters.

Integrating SVN with hSBM

This section illustrates how SVN is integrated with hSBM by using SVN as a filtering tool on the bipartite graph. SVN identifies statistically significant word co-occurrences and discards irrelevant connections, improving the quality of the graph.

Once the bipartite graph is filtered by SVN, it becomes a more reliable representation of document-word relationships, making it suitable for subsequent community detection tasks using hSBM. We compare two approaches for building the bipartite graph. We replace the set of nodes with: i) the word-pairs, as illustrated in Fig. 4, ii) the set of words that were statistically validated in Step 1.

Table 2 presents the document clustering performance of combining SVN and hSBM, excluding the hyperlink layer. We consider SVN methods with α values of 0.01 and 0.05 denoted as SVN_1 and SVN_5, respectively. Instances where SVN and hSBM are combined are indicated by the "+" symbol, signifying the application of hSBM after applying the SVN method. The second column specifies the layers of the bipartite network included in the model, while the third and fourth columns present performance scores for the two measures under consideration. Finally, the fifth column reports the models'

Table 2 MPO and NMI scores between the models and the true partition. The last column reports the Description Length. The values correspond to the mean over 100 replicates; standard deviation in parenthesis

Models	Layers	MPO (Peixoto 2021)	NMI (Fano 1961)	DL (Grünwald 2007)
Wikipedia 120				
hSBM	Docs, Words	0.59 (0.10)	0.35 (0.12)	229 457(1194)
SVN_1 + hSBM	Docs, WordPairs	0.62 (0.09)	0.38 (0.04)	87 104(369)
SVN_5 + hSBM	Docs, WordPairs	0.64 (0.10)	0.40 (0.04)	103 512(360)
SVN_1 + hSBM	Docs, Words	0.56 (0.10)	0.36 (0.04)	105 208(554)
SVN_5 + hSBM	Docs, Words	0.54 (0.11)	0.36 (0.06)	114 693(595)
arXiv 10k				
hSBM	Docs, Words	0.43 (0.03)	0.39 (0.04)	4 590 912 (8656)
SVN_1 + hSBM	Docs, WordPairs	0.52 (0.02)	0.50 (0.02)	2 853 771 (2 923)
SVN_5 + hSBM	Docs, WordPairs	0.54 (0.01)	0.51 (0.01)	3 130 354 (1 726)
SVN_1 + hSBM	Docs, Words	0.52 (0.04)	0.50 (0.02)	2 795 090 (1 161)
SVN_5 + hSBM	Docs, Words	0.53 (0.02)	0.51 (0.02)	2 852 855 (1 509)
arXiv 100k				
hSBM	Docs, Words	0.44 (0.06)	0.37 (0.03)	34 697 318 (49 200)
SVN_1 + hSBM	Docs, WordPairs	0.39 (0.02)	0.31 (0.01)	30 443 765 (38 665)
SVN_5 + hSBM	Docs, WordPairs	0.39 (0.01)	0.32 (0.02)	32 803 636 (11 250)
SVN_1 + hSBM	Docs, Words	0.37 (0.02)	0.32 (0.01)	26 552 293 (8 104)
SVN_5 + hSBM	Docs, Words	0.32 (0.02)	0.30 (0.02)	28 726 302 (9 650)

performance in terms of Description Length (Rissanen 1978; Grünwald 2007), a metric indicating the model's ability to compress essential information for describing both the data and the model parameters.

The results show that integrating SVN with hSBM leads to a slight performance improvement in both metrics compared to hSBM alone. For instance, in the *Wikipedia 120* dataset, the MPO for hSBM with the word layer is 0.59 ± 0.10, while SVN_1 + hSBM (using word-pairs) achieves a MPO of 0.62 ± 0.09, and SVN_5 + hSBM achieves 0.64 ± 0.10. Similarly, NMI scores also show slight improvements: hSBM with words yields 0.35 ± 0.12, while SVN_1 + hSBM with word-pairs yields 0.38 ± 0.04 and SVN_5 + hSBM with word-pairs reaches 0.40 ± 0.04.

This is particularly significant because the filtered bipartite graphs, which are reduced in size (fewer nodes), result in lower Description Length (DL) values, reflecting a more compact representation of the information. Despite this reduction, the performance is improved. For example, in the *Wikipedia 120* dataset, the DL for the original word layer is $229\,457 \pm 1{,}194$, while for SVN_1 + hSBM with word-pairs, the DL is significantly reduced to $87\,104 \pm 369$. This confirms that the SVN method effectively retains the most important word-pair relationships, even with fewer nodes and a more compact graph representation.

Finally, the results reveal that the word-pairs layer provides superior performance on the *Wikipedia 120* dataset, with a notable increase in MPO and NMI scores when compared to the single word layer.

The performance advantages are particularly pronounced in the *arXiv 10k* dataset, where the integration of SVN leads to significant improvements over the baseline hSBM. For example, hSBM with the original word layer achieves an MPO of 0.43 ± 0.03 and an NMI of 0.39 ± 0.04, whereas SVN_1 and SVN_5 with word-pair layers achieve MPO scores of 0.52 ± 0.02 and 0.54 ± 0.01, respectively, and NMI scores of 0.50 ± 0.02 and 0.51 ± 0.01. Similarly, the reduction in Description Length (DL) for SVN-based methods is substantial, with SVN_1 reducing the DL from approximately 4, 590, 912 to 2, 853, 771.

However, in the *arXiv 100k* dataset, hSBM with the original word layer performs better overall compared to the SVN-integrated configurations. While SVN_1 and SVN_5 still achieve a more compact representation with lower DL values, their MPO and NMI scores (e.g., 0.39 ± 0.02 and 0.31 ± 0.01, respectively, for SVN_5) fall short of those achieved by the baseline hSBM, which records an MPO of 0.44 ± 0.06 and an NMI of 0.37 ± 0.03. This indicates that for larger datasets like *arXiv 100k*, the benefits of reducing the set of nodes related to words might diminish.

Topic modeling

In this subsection, we present the outcomes concerning the task of topic modeling, which involves Step 3 of our methodology, and compare our methods with hSBM, LDA, and BERTopic_refit (the refitted version). Among the two models proposed, WCSVNtm_1 and WCSVNtm_5, we report the results for WCSVNtm_5 since it shows highest scores in performing document clustering in all the three datasets.

We evaluate the performance of several topic models on three different metrics: Coh_{SVN} (Simonetti et al. 2023), CV, and PMI (pointwise mutual information) (Röder

et al. 2015) across the three aforementioned datasets: *Wikipedia 120*, *arXiv 10k*, and *arXiv 100k*. A detailed comparison of the models is shown in Table 3. As mentioned before, for the hSBM there is not a "natural" threshold for cutting the hierarchical tree of topics. Therefore, we selected the first two levels from the root. For BERTopic, to prevent over-fragmentation we consider the refitted version of the model, fixing the number of topics to match those discovered by WCSVNtm.

On the smallest dataset, BERTopic_refit achieves the highest Coh_{SVN} (0.46) and CV (0.70) scores, along with the best PMI (0.26). However, this comes with a significant limitation: the model identifies only two topics across 120 documents. While these two topics exhibit high coherence, the small number of topics identified is problematic for capturing the full diversity of the dataset. This issue highlights the trade-off between coherence and granularity-while the model produces highly coherent topics, it fails to adequately represent the variety of content across the documents.

In contrast, WCSVNtm achieves a better balance between coherence and diversity, with Coh_{SVN} (0.36) and CV (0.65) scores that are slightly lower than BERTopic's, but with 20 topics identified, providing a richer representation of the dataset.

The hSBM models, particularly those integrating hyperlink data (e.g., *hSBM (35) - Text+Hyper*), perform poorly on all metrics, with Coh_{SVN} and CV scores under 0.15 and PMI values as low as −5.39. These results suggest that the inclusion of hyperlink information does not enhance, and may even detract from, the model's ability to generate coherent topics.

Table 3 Comparison of topic models through topic coherence scores Coh_{SVN}, CV, and PMI metrics for the Wikipedia and arXiv datasets. The number of topics identified by each model is reported in parentheses

Model	Coh_{SVN}	CV	PMI
Wikipedia 120			
LDA (20)	0.25	0.54	− 1.73
WCSVNtm5(20)	0.36	0.65	− 2.13
hSBM (64) - Text	0.13	0.44	− 6.11
hSBM (11) - Text	0.14	0.47	− 1.47
hSBM (35) - Text+Hyper	0.08	0.43	− 5.39
hSBM (8) - Text+Hyper	0.09	0.43	− 1.61
hSBM+SVN5 (10) - Text	0.07	0.47	− 12.74
BERTopic (2)	0.46	0.70	0.26
arXiv 10k			
LDA (47)	0.13	0.39	− 5.12
WCSVNtm_5 (47)	0.33	0.58	− 0.74
hSBM (59) - Text	0.17	0.48	− 2.82
hSBM (8) - Text	0.22	0.51	0.11
BERTopic_refit (47)	0.25	0.47	− 2.74
arXiv 100k			
LDA (108)	0.13	0.33	− 0.41
WCSVNtm_5 (108)	0.26	0.47	0.04
hSBM (103) - Text	0.20	0.43	− 3.26
hSBM (16) - Text	0.16	0.35	− 2.91
BERTopic_refit (108)	0.40	0.58	− 0.28

On the *arXiv 10k* dataset, WCSVNtm stands out with the highest Coh*SVN* score (0.33) and CV score (0.58), underscoring its robustness in identifying coherent topics. Although its PMI score (−0.74) is lower than that of *hSBM (8)* (0.11). In contrast, BERTopic_refit and *LDA* perform less effectively, with their Coh*SVN* and CV scores lagging behind those of WCSVNtm, highlighting their limited ability to capture meaningful topics in medium-sized datasets like *arXiv 10k*. Interestingly, *hSBM (8)*, despite its smaller number of topics, achieves a positive PMI score (0.11), suggesting that this configuration captures some meaningful relationships between terms. However, its Coh_{SVN} and CV scores remain moderate, reflecting a less cohesive clustering approach compared to WCSVNtm.

For the largest dataset, *arXiv 100k*, WCSVNtm delivers strong performance, achieving the highest PMI score (0.04) and competitive Coh_{SVN} (0.26) and CV (0.47) scores, demonstrating its ability to generate coherent and meaningful topics at such large scale. The hSBM models generally underperform on the *arXiv 100k* dataset, with low scores across all metrics. The smaller configuration, *hSBM (16)*, shows marginally better PMI (−2.91) but struggles with coherence. BERTopic_refit, after addressing the over-fragmentation issue by refitting the model and fixing the number of topics to match those of WCSVNtm (108 topics), achieves the highest CV score (0.58) and an impressive Coh_{SVN} (0.40). However, its PMI score remains slightly negative (−0.28).

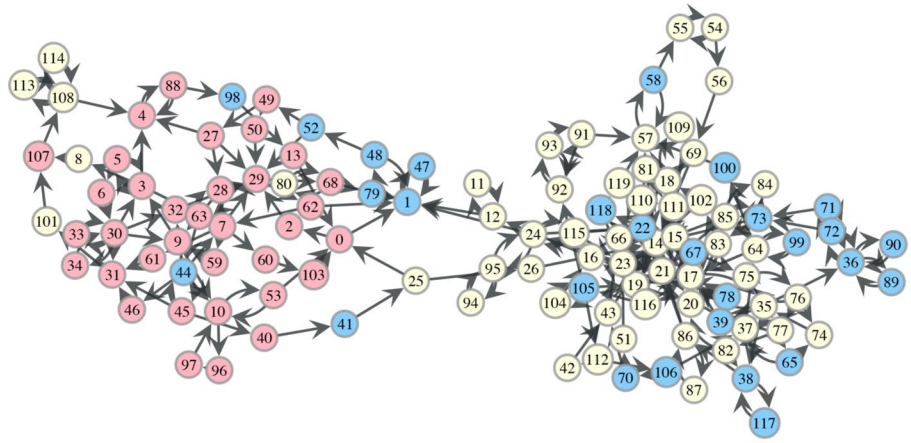
Interpretation of document partitions and topics

In this section, we present various visualization techniques that can aid in interpreting document partitions, topics, and topic-document associations. Given the size of the dataset, we focus on the results from the Wikipedia dataset, as the relatively small number of 120 articles provides a clearer and more interpretable visualization. We extract topics from the validated word co-occurrence network, using the Leiden community detection algorithm (Blondel et al. 2008), which results in 20 distinct topics. Then, we assign topics to documents by evaluating the overlap between the words within a community (topic) and those within a document (Step 4). We conduct the Fisher's exact test by counting the words shared by a document and a topic, along with their marginal counts. The test considers only the words in the resulting statistically validated network as the universe, as depicted in Fig. 5. Subsequently, for each topic, we evaluate the over-expression for each document by assessing the number of documents that share more than one word with the topic. We set $\alpha = 0.05$ when employing the False Discovery Rate (Benjamini and Hochberg 1995).

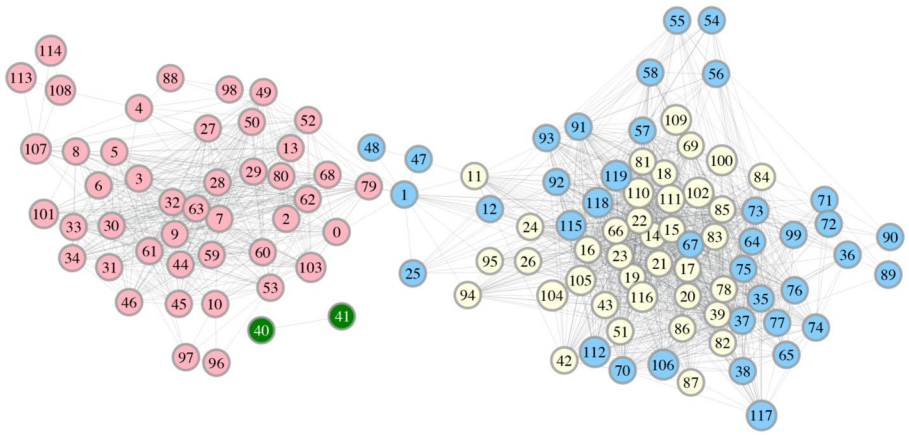
Network visualizations

In Fig. 7, we visually compare the partition of the original network with the community structure retrieved from the statistically validated network of documents. Figure 7a depicts the original network with the true partition in scientific categories (Biology in pink, Physics in blue, Maths in yellow). While Fig. 7b displays the SVN of documents and the retrieved community structure.

The visual comparison of networks highlights the effectiveness of our approach in representing documents as nodes in a network. This is achieved through a two-step



(a) *Network of hyperlinks with original Wikipedia partition.*



(b) *SVN of documents with Leiden partition.*

Fig. 7 Comparing Network Partitions: (**a**) original Wikipedia partition, (**b**) SVN projection on the set of documents and Leiden partition. In the two network representations, all nodes are placed in the same positions

filtering procedure that eliminates random (according to the null hypothesis) word co-occurrences within sentences and subsequently considers only the statistically significant overlap of documents based on shared vocabulary.

Comparing Fig. 7a and b, we note that some topological structures are present in both networks. The nodes related to Biology (pink) are well intra-connected and separated from Mathematics (yellow) and Physics (blue) in both networks. Furthermore, the same nodes bridge the two communities in both networks—nodes (ID 1, 47 and 48) in Fig. 7b (representing the interdisciplinary documents), the group of nodes on the top left of the figure (ID 113, 114 and 108).

Document-topic association

To gain deeper insights into the topics identified in the Wikipedia dataset, we first present them in Table 4. The most representative words within each topic were identified using the importance score based on modularity. Additionally, each of the 20 topics was assigned a descriptive label through a topic labeling process, using ChatGPT to generate names based on the most representative words.

The Sankey diagram in Fig. 8 illustrates the extent to which the 20 topics are over-represented in the original partitions, as well as in the partitions extracted from the SVN of documents, consisting of the 4 distinct communities (already represented in Fig. 7).

The Sankey diagram allows us to visualize the number of documents within each community that exhibit over-expression of specific topics. Notably, a single topic can be over-expressed across different documents, and some documents may show over-expression of multiple topics.

From the Sankey diagram we note that almost all Biology documents are grouped within community Pink, reflecting a strong and cohesive relationship between this tag and topics like *Cellular and Molecular Biology*, *Epigenetics and Gene Regulation*, and *Transcription Factors and Gene Expression*. However, Biology's influence extends beyond its primary community.

A few Physics documents exhibit associations with Biology topics, bridging distinct communities. These include documents named as *X-ray Crystallography* (ID 1), *Direct Methods (Crystallography)* (ID 47), and *Multipole Density Formalism* (ID 48), which sit

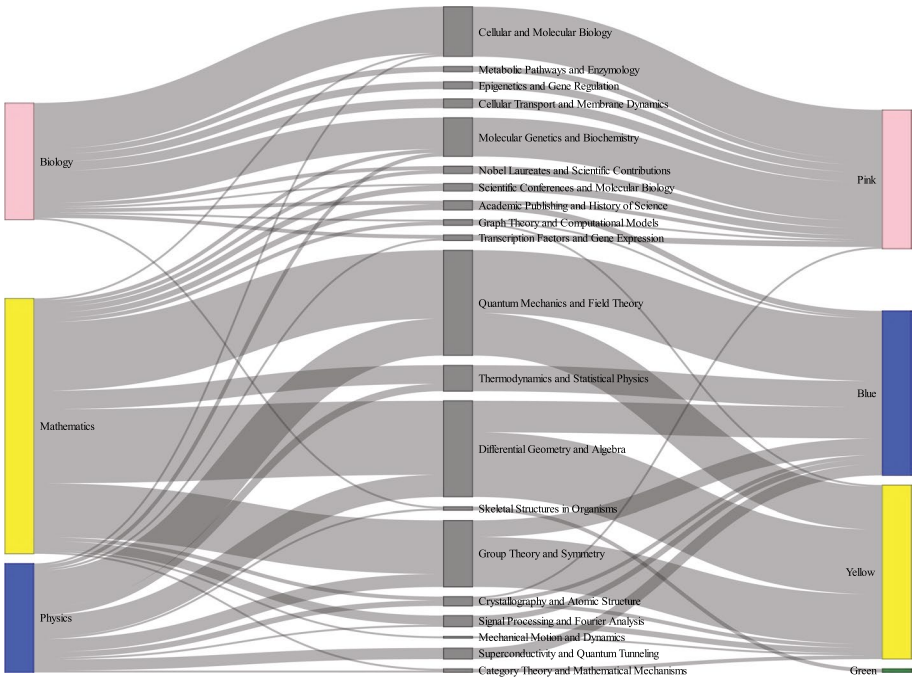


Fig. 8 Sankey diagram showing the flow from Wikipedia article tags (Biology, Mathematics, Physics) to topics identified by our method (middle) and their distribution across document communities (Pink, Blue, Green, Yellow). Line thickness indicates the number of documents in each group (on left and right panels) over-expressing a topic. The size of each node is determined by the maximum between the incoming and outgoing flows

Table 4 Topic labels, their corresponding IDs, and representative words extracted from the validated word co-occurrence network. The labels were generated using ChatGPT

Topic id	Label	Words
1	Cellular and Molecular Biology	extrins signal canin virus caspas activ opioid pain factor promot cycl regul break embryon mous control cell transcript dna bind
2	Quantum Mechanics and Field Theory	negat posit perturb theori hamiltonian field quantum motion topolog classic quantize goldston scale wilson metric spacetim boson flat equat hamilton
3	Molecular Genetics and Biochemistry	translat acid sequenc structur bond coval protein exon mrna ankyrin fold genet materi snrna transfer complementari polyadenyl genom groov helix
4	Differential Geometry and Algebra	matrix space lie defin pullback algebra homogen princip isomorph tangent exterior map differenti algebroid bracket column row ani product subspac
5	Group Theory and Symmetry	represent angular axi charact tabl diatom rotat left right element subgroup invari heteronuclear haar set vibron local dipol compact boost
6	Metabolic Pathways and Enzymology	metabol phosphat cytochrom ferment lactic enzym pathway acetyl mitochondrion glycogen starch bax proapoptot concentr phosphofructokinas step yeast gluconeogenesi glucos bak
7	Crystallography and Atomic Structure	diffract cubic hexagon detect radioact atom crystal chemic isotop ratio determin mmm orthorhomb zeta triclin tetragon trigon beam monochromat intens
8	Thermodynamics and Statistical Physics	energi state surround temperatur thermodynam volum system variabl evolut heat effici reservoir law clausius entropi revers intern isol union time
9	Superconductivity and Quantum Tunneling	junction tunnel josephson phase appli current insul flux penetr abrikosov superconductor expon phenomenon eq barrier vortex voltag superconduct magnet microscal
10	Signal Processing and Fourier Analysis	algorithm fft discret fast transform fourier seri dtft sampl legendr comb dft dimens primari librari rotam conform eigenvector alias interpol
11	Academic Publishing and History of Science	survey york american societi son wiley cambridg uk postscript seminar hodgkin vincent rivasseau text gold green new london avail crick
12	Epigenetics and Gene Regulation	skew symptom repress matern patern express gene inact xi chromosom heterozyg cluster parahox xist drosophila inactiv normal epigenet random coat
13	Graph Theory and Computational Models	configur vertex grain unstabl graph complex assess benchmark grid small recurr sandpil antenna dock stabl rectangular reach model collaps sand
14	Nobel Laureates and Scientific Contributions	lesli orgel nobel sulston physiolog brenner award prize john sydney cormack greek earth life institut horvitz medicin california diego research
15	Scientific Conferences and Molecular Biology	si unit ismb scientif confer intellig organis europ hold dogma molecular talk track joint biolog meet venu comput european bioinformat
16	Skeletal Structures in Organisms	endoskeleton support bacteri plant miner organ bone cartilag pelvi skeleton compos fish exoskeleton shell fossil tough clam anim skelet develop
17	Cellular Transport and Membrane Dynamics	cytoplasm envelop membran export nuclear nucleus organell cargo exportin memori neuron bodi transport red nucleolus inner rangtp importin lamina pore
18	Transcription Factors and Gene Expression	tfiid tfiie box tbp tfiib tfiif preiniti tfb tfiia tfiih rnap archaeal tata
19	Mechanical Motion and Dynamics	clockwis fh fd id fv horizont fc counterclockwis middl
20	Category Theory and Mathematical Morphisms	hbordm morphism categori tqft bordism object functor

https://chat.openai.com/chat

at the boundary between Physics and Biology (see Fig. 7 to visualize them in the network). Furthermore, Physics documents fully embedded within the Pink community, such as *Isotopic Labeling* (ID 98), *Transcription Factor* (ID 44), and *Macromolecule* (ID 52) (see Fig. 7), demonstrate the interdisciplinary reach of Biology topics into Physics.

Mathematics documents also show intersections with Biology, particularly in topics related to computational and applied biology. Documents such as *Bioinformatics Open Source Conference* (ID 114), *European Conference on Computational Biology* (ID 113), and *Intelligent Systems for Molecular Biology* (ID 108) (which can be identified in the top-left region of Fig. 7) establish a clear connection between Mathematics and Biology, especially in areas focused on conferences and computational modeling. Additionally, Biology-related Mathematics documents embedded within the Pink community, such as *The Proteolysis Map* (ID 8), *TopFIND* (ID 101), and *Modeling Biological Systems* (ID 80), highlight cross-disciplinary influences.

While Biology dominates the Pink community, Physics and Mathematics exhibit a significant overlap in topics, as reflected in their connections with the Blue and Yellow communities. Topics such as *Quantum Mechanics and Field Theory*, *Differential Geometry and Algebra*, and *Group Theory and Symmetry* are strongly represented in documents spanning both fields. This overlap underscores the connections between Mathematics and Physics. The Blue community, in particular, captures topics unique to Physics, such as *Thermodynamics and Statistical Physics*, *Skeletal Structures in Organisms*, and other subfields, illustrating how Physics retains its distinct identity while sharing space with Mathematics.

The Green partition stands out as a small community, consisting of just two documents: one from Physics (ID 41, Exoskeleton) and one from Biology (ID 40, Skeleton). This community focuses on *Skeletal Structures in Organisms* reflecting a meaningful intersection between the two disciplines.

The Sankey diagram ultimately highlights the structure of interdisciplinarity within this dataset. Through the double projection on words and documents, we achieve a more detailed description of the dataset. Even though Mathematics and Physics share substantial overlap, the method still manages to group the documents well.

Conclusions

In this study, we introduced the Word Co-occurrence SVN Topic Model (WCSVNtm), a novel approach for document clustering and topic modeling that leverages a Statistically Validated Network (SVN) framework to extract significant word co-occurrences and thematic document groupings, addressing key challenges in analysing textual data as bipartite networks. Our method aligns with recent advancements in network-based approaches (Gerlach et al. 2018; Hyland et al. 2021; Veremyev et al. 2019), offering a robust framework for uncovering semantic relationships in textual datasets.

We validated our method on three datasets: a set of 120 Wikipedia articles, the arXiv10 dataset, which consists of 100 000 abstracts from scientific papers, and a sampled subset of 10 000 documents from the original arXiv10. In both document clustering and topic modeling, WCSVNtm demonstrated competitive performance compared to state-of-the-art methods, such as hSBM, BERTopic, and LDA, while maintaining computational efficiency, reproducibility, and robustness. However, it

is important to highlight certain considerations about the comparison among methods. For hSBM, deriving a partition of the corpus is not straightforward, as it requires carefully selecting the hierarchical levels where the tree should be cut to retrieve clusters. In our analysis, this selection was guided by prior knowledge of the "true" partition, optimizing hSBM's results. However, this reliance on prior knowledge is impractical in real-world unsupervised scenarios, where such information is typically unavailable. Similarly, BERTopic exhibited significant over-fragmentation in the *arXiv10k* and *arXiv100k* datasets, initially producing an excessively large number of topics and document clusters. To address this issue, we performed a refitting step, fixing the number of topics to align with WCSVNtm's results. While this adjustment improved the computed metrics, it highlights a lack of out-of-the-box robustness when handling large-scale datasets in an unsupervised context, especially in the document clustering task. Additionally, on the smaller Wikipedia dataset, BERTopic struggled to represent the variety of content, identifying only two clusters of documents and corresponding topics.

While existing state-of-the-art probabilistic techniques (e.g. hSBM, LDA) primarily rely on analysing document-term matrices, our approach stands out by focusing on sentence-level co-occurrences, enabling a more detailed examination of the local context of words. Furthermore, our use of a word-pair-document network, rather than a simpler word-document network, provides a richer representation of relationships between words. Unlike word-document networks, which only capture the presence of words in documents, the word-pair-document network models co-occurrence relationships, preserving valuable contextual information. This approach facilitates the understanding of how words are related. Words that frequently appear together in the same context are more likely to be semantically related, and the word-pair-document network allows us to model these semantic relationships more effectively. As a result, this richer network representation provides more detailed information, improving the identification of topics and clusters of documents.

A key innovation of our method is the introduction of rigorous statistical tests to filter out spurious relationships among words, to evaluate similarity between documents, and establish a threshold to test the association between topics and documents. Building upon the work of Simonetti et al. (2023), which demonstrated the power of SVN methods in uncovering statistically significant word connections, our extension focuses on directly extracting topics and leveraging semantic relationships for document clustering. By prioritizing word-pairs over single-word occurrences, our approach captures complex semantic information more effectively, representing a significant advancement in network-based text analysis. Additionally, the results from integrating SVN with hSBM highlight the effectiveness of the introduced test statistics in identifying significant word co-occurrences and eliminating irrelevant connections among words.

In conclusion, the application of SVN to textual data represents a powerful tool for unsupervised extraction of fundamental distributional semantic relations among words. Specifically, the proposed method allows the automatic selection of both the number of topics and document clusters through the Leiden community detection algorithm. This is a valuable feature, particularly in unsupervised settings, which

enhances the proposed methodology compared to other state-of-the-art methods. Furthermore, our method can be further improved as the research in network science progresses in exploiting cutting-edge algorithms for community detection.

Future work could also explore alternative corrections for multiple testing beyond Bonferroni and FDR, potentially enhancing the sensitivity and robustness of the network construction process. Our approach holds promise for advancing how network science tools can effectively explore semantic analysis in textual data.

Appendix 1: Stochastic block models

Stochastic Block models (SBMs) are a class of random graph models that generate networks with an adjacency matrix A_{ij} based on the probability $P(\boldsymbol{A} \mid \boldsymbol{b})$, where the vector $\boldsymbol{b}$ with entries $b_i \in \{1, \ldots, B\}$ indicates the group membership of node $i = 1, \ldots, D$ among B possible groups. For the multilayer network design we developed-addressing the three types of data (H, T, M)-we apply the SBM framework to each individual layer and combine them by ensuring that document groups are consistent across all layers, resulting in a joint probability:

$$P(\boldsymbol{A}_H, \boldsymbol{A}_T, \boldsymbol{A}_M \mid \boldsymbol{b}) = P(\boldsymbol{A}_H \mid b)P(\boldsymbol{A}_T \mid \boldsymbol{b})P(\boldsymbol{A}_M \mid \boldsymbol{b}),$$

where $\boldsymbol{A}_H$, $\boldsymbol{A}_T$, and $\boldsymbol{A}_M$ represent the adjacency matrices for each respective layer. In each layer, edges between nodes i and j are drawn from a Poisson distribution with mean (Karrer and Newman 2011):

$$\theta_i \theta_j \omega_{b_i, b_j},$$

where ω_{rs} is the expected number of edges between group r and group s, b_i denotes the group membership of node i, and θ_i is the propensity for node i to be selected within its own group. Non-informative priors are used for the parameters θ and ω, and the marginal likelihood of the SBM is calculated as (Peixoto 2017):

$$P(\boldsymbol{A} \mid \boldsymbol{b}) = \int P(\boldsymbol{A} \mid \boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{b}) P(\boldsymbol{\omega}, \boldsymbol{\theta} \mid \boldsymbol{b})\, d\boldsymbol{\theta}\, d\boldsymbol{\omega}.$$

From this, we obtain the posterior distribution for a single partition conditioned on the edges from all layers (Hric et al. 2014):

$$P(\boldsymbol{b} \mid \boldsymbol{A}_H, \boldsymbol{A}_T, \boldsymbol{A}_M) = \frac{P(\boldsymbol{A}_H \mid \boldsymbol{b})P(\boldsymbol{A}_T \mid \boldsymbol{b})P(\boldsymbol{A}_M \mid \boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{A}_H, \boldsymbol{A}_T, \boldsymbol{A}_M)}.$$

This method enables the clustering of both words and documents into categories.

Appendix 2: Evaluation metrics

1.1. Normalized mutual information (NMI)

Normalized Mutual Information (NMI) is a metric commonly used to assess the similarity between two partitions, considering the mutual information between the partitions while normalizing for the total entropy of the partitions. NMI is particularly useful in

comparing partitions when the number of clusters or communities differs between the partitions.

Given two partitions $\boldsymbol{x}$ and $\boldsymbol{y}$ of a set of nodes, the mutual information $I(\boldsymbol{x}, \boldsymbol{y})$ between these partitions is defined as:

$$I(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \sum_j p(x_i, y_j) \log \left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)} \right),$$

where $p(x_i)$ and $p(y_j)$ are the marginal probabilities of nodes in partitions $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, and $p(x_i, y_j)$ is the joint probability that node i belongs to cluster x_i in partition $\boldsymbol{x}$ and cluster y_j in partition $\boldsymbol{y}$.

The entropy of a partition $\boldsymbol{x}$ is defined as:

$$H(\boldsymbol{x}) = -\sum_i p(x_i) \log p(x_i),$$

and similarly for $\boldsymbol{y}$:

$$H(\boldsymbol{y}) = -\sum_j p(y_j) \log p(y_j).$$

The Normalized Mutual Information (NMI) is then given by:

$$\mathrm{NMI}(\boldsymbol{x}, \boldsymbol{y}) = \frac{2I(\boldsymbol{x}, \boldsymbol{y})}{H(\boldsymbol{x}) + H(\boldsymbol{y})},$$

where the factor of 2 ensures that NMI lies between 0 and 1. The NMI is 1 when the two partitions are identical (i.e., they have maximal mutual information), and 0 when there is no mutual information between the partitions.

1.2. Maximum partition overlap

Maximum Partition Overlap (MPO), proposed by Peixoto (2021), is a metric used to evaluate how much overlap exists between two community partitions, particularly when nodes can belong to multiple communities.

The maximum overlap between partitions $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by:

$$w(\boldsymbol{x}, \boldsymbol{y}) = \arg\max_{\mu} \sum_i \delta_{x_i, \mu(y_i)},$$

where $\boldsymbol{\mu}$ is a bijective mapping between the group labels.

The normalized maximum overlap between partitions $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as:

$$w(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{N} \sum_i \delta_{x_i, y_i},$$

where N is the number of nodes, and $w(x, y)$ lies in the unit interval $[0, 1]$.

Given multiple partitions, we also seek to extract a consensus partition $\hat{\boldsymbol{b}}$, which maximizes the sum of overlaps with all partitions. This consensus partition can be obtained by double maximization of the following set of equations:

$$\hat{b}_i = \arg\max_r \sum_m \delta_{\mu_m(b_i^m),r},$$

$$\boldsymbol{\mu}_m = \arg\max_{\boldsymbol{\mu}} \sum_r m^{(m)}_{r,\mu(r)},$$

where $\boldsymbol{\mu}$ is a bijective mapping between the group labels, and $m^{(m)}_{r,\mu(r)}$ represents the contingency table between $\hat{\boldsymbol{b}}$ and partition $\boldsymbol{b}^{(m)}$.

An iterative procedure is then applied to this set of equations until no further improvement is possible. The uncertainty σ of the consensus partition obtained from M_p partitions is quantified as (Peixoto 2021):

$$\sigma = 1 - \frac{1}{NM_p} \sum_i \sum_m \delta_{\mu_m(b_i^m),\hat{b}_i}.$$

The maximum partition overlap is 1 when the two partitions are identical in terms of community membership, and 0 when there is no overlap.

1.3. Description length

The Description Length (DL) is used to measure the *complexity* of a given community structure. It quantifies how well the structure captures the underlying data, with the goal of identifying the simplest partition that best explains the observed patterns in the system. Essentially, the description length assesses the efficiency of a model in representing the data.

The DL is defined as:

$$\mathrm{DL} = -\log P(\boldsymbol{A}_H, \boldsymbol{A}_T, \boldsymbol{A}_M, b),$$

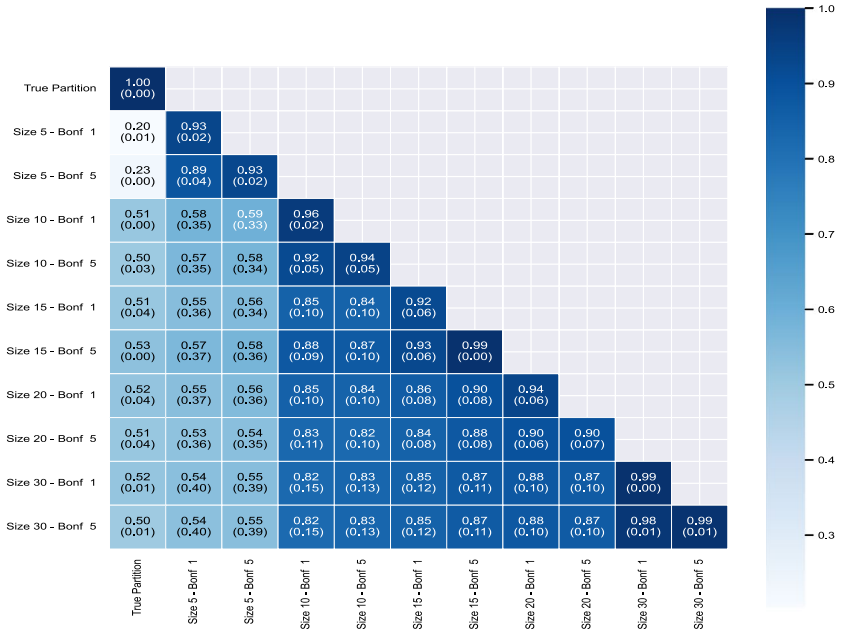
where $\boldsymbol{A}_H$, $\boldsymbol{A}_T$, and $\boldsymbol{A}_M$ represent the adjacency matrices for the respective layers of the model, and b represents the community partition.

In practical terms, a lower description length indicates that the model is more compact and efficient in explaining the data.

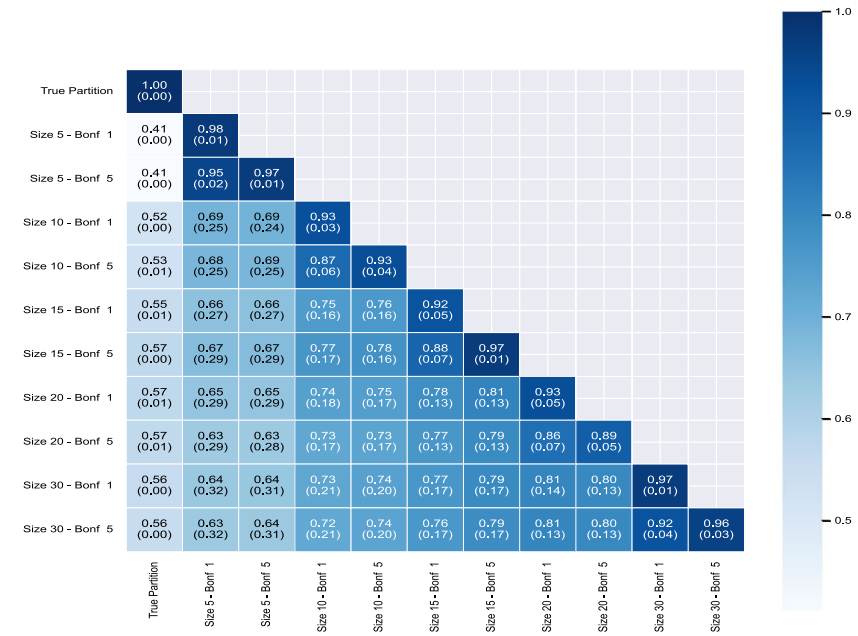
Appendix 3: Sensitivity analysis on arXiv 10k

We have performed a train-test split stratified by subject (70%-30%) on the 10k article sample of the arXiv dataset. While the optimization of the length of the sliding window is not strictly necessary in this case—since the arXiv dataset contains original sentences, and we can directly compute co-occurrences without the need for a sliding window—we implemented it to study the sensitivity of the model to hyperparameters on a large-scale dataset and to demonstrate its robustness.

The sensitivity analysis (Heatmaps in Fig. 9a) on the training set (in sample) confirms that both NMI and MPO metrics remain relatively stable across a range of hyperparameter values, except for a window size of 5, which produces suboptimal results. For window sizes from 10 upward, the results vary only slightly (e.g., MPO ranges from 0.51 to 0.53, and NMI ranges from 0.52 to 0.58).

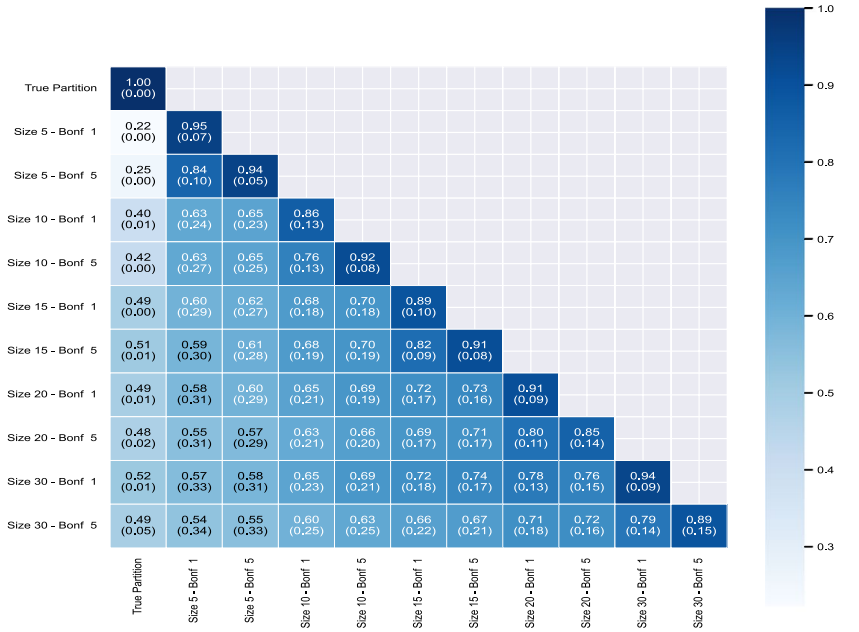


(a) *Maximum partition overlap measure.*

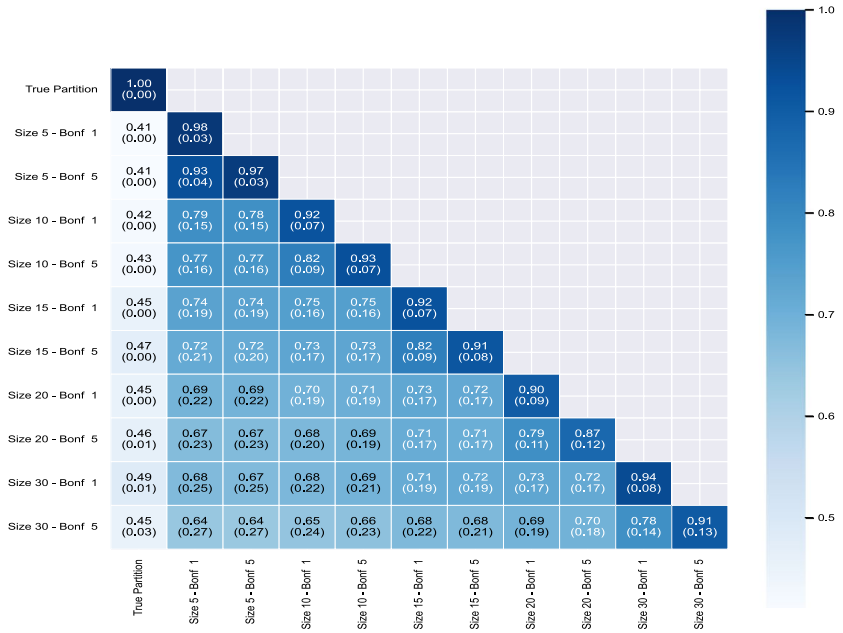


(b) *Normalized Mutual Information (NMI) measure.*

Fig. 9 Heatmaps. Dataset: arXiv10 Sub sample of 10 000, Train set (70%). The average and standard deviation (in parenthesis) among 100 replicates using the MPO (**a**) and NMI (**b**) between and within different parameter settings. The scores represent the average over 100 replicates of applying the Leiden community detection algorithm on the SVN of documents to extract the document partition



(a) *Maximum partition overlap measure.*



(b) *Normalized Mutual Information (NMI) measure.*

Fig. 10 Heatmaps. Dataset: arXiv10 Sub sample of 10 000, Test set (30%). The average and standard deviation (in parenthesis) among 100 replicates using the MPO (**a**) and NMI (**b**) between and within different parameter settings. The scores represent the average over 100 replicates of applying the Leiden community detection algorithm on the SVN of documents to extract the document partition

As for the out of sample analysis, we present the same heatmap obtained from the Test set in Fig. 10. The results remain stable also in the test set. Moreover, if the optimal values of α and window size are selected from the training set ($\alpha = 0.05$ and window size equal 15), the results from the test set turn out to be the second best (according to MPO) and withing a standard deviation from the best one. This result, together with the stability of performance at varying values fo window size and threshold suggest that the method is not prone to overfitting.

Acknowledgements
We would like to express our gratitude to the reviewers for their valuable feedback and constructive suggestions, which have significantly improved the quality of this work.

Author contributions
AS: conceptualization, software, formal analysis, data curation, writing - original draft, visualization. AA: conceptualization, writing - original draft, visualization. MT: methodology, formal analysis, writing - review and editing, supervision. TDM: methodology, writing - review and editing, supervision.

Funding
The reserach work of Andrea Simonetti and Michele Tumminello was partially funded by the European Union - NextGenerationEU, in the framework of the GRINS -Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 - CUP B73C22001260006). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them. The research work of Alessandro Albano has been supported by the European Union - NextGenerationEU - National Sustainable Mobility Center CN00000023, Italian Ministry of University and Research Decree n. 1033- 17/06/2022, Spoke 2, CUP B73C2200076000.

Data availability
The first dataset analysed during the current study is provided by Hyland et al. (2021), and it is available at the following link: https://github.com/martingerlach/hSBMTopicmodel/tree/master/data.
The second dataset used in the analysis is available at https://paperswithcode.com/dataset/arxiv-10.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 30 April 2024 Accepted: 29 January 2025
Published online: 19 February 2025

References

Airoldi EM, Blei DM, Fienberg SE, Xing EP (2008) Mixed membership stochastic Blockmodels. J Mach Learn Res 9:1981–2014
Altuncu MT, Mayer E, Yaliraki SN, Barahona M (2019) From free text to clusters of content in health records: an unsupervised graph partitioning approach. Appl Netw Sci 4:1–23
Altuncu MT, Yaliraki SN, Barahona M (2021) Graph-based topic extraction from vector embeddings of text documents: application to a corpus of news articles. Complex networks & their applications IX: Volume 2, proceedings of the ninth international conference on complex networks and their applications complex networks 2020 (pp 154–166)
Angelov D (2020) Top2vec: distributed representations of topics. arXiv:2008.09470
Ball B, Karrer B, Newman MEJ (2011) Efficient and principled method for detecting communities in networks. Phys Rev E 84:036103
Belford M, Mac Namee B, Greene D (2018) Stability of topic modeling via matrix factorization. Expert Syst Appl 91:159–169
Bengio Y, Ducharme R, Vincent P, Janvin C (2003) A neural probabilistic language model. J Mach Learn Res 3:1137–1155
Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodol) 57(1):289–300
Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 1165–1188
Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res3:993–1022
Blondel VD, Guillaume J- L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008
Casella G, Berger RL (2002) Statistical inference. Duxbury Press
Dieng AB, Ruiz FJ, Blei DM (2020) Topic modeling in embedding spaces. Trans Assoc Comput Ling 8:439–453
Fano RM (1961) Transmission of information: a statistical theory of communications. MIT Press, Cambridge, MA

Farhangi A, Sui N, Hua N, Bai H, Huang A, Guo Z (2022) Protoformer: embedding prototypes for transformers. Advances in knowledge discovery and data mining: 26th pacific-asia conference, Pakdd 2022, Chengdu, China, May 16–19, 2022, proceedings, part i, pp 447–458
Fortunato S (2010) Community detection in graphs. Phys Rep 486(3–5):75–174
Gerlach M, Peixoto TP, Altmann EG (2018) A network approach to topic models. Sci Adv 4(7):eaaq1360
Gramatica R, Di Matteo T, Giorgetti S, Barbiani M, Bevec D, Aste T (2014) Graph theory enables drug repurposing-how a mathematical model can drive the discovery of hidden mechanisms of action. PLoS ONE 9(1):e84912
Grootendorst M (2022) Bertopic: neural topic modeling with a class-based tf-idf procedure. arXiv:2203.05794
Grünwald PD (2007) The minimum description length principle. MIT Press
Gururangan S, Marasović A, Swayamdipta S, o K, Beltagy I, Downey D, Smith NA (2020) Don't stop pretraining: adapt language models to domains and tasks. arXiv:2004.10964
Hatzopoulos V, Iori G, Mantegna RN, Micciche S, Tumminello M (2015) Quantifying preferential trading in the e-mid interbank market. Quant Finance 15(4):693–710
Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: first steps. Soc Netw 5(2):109–137
Hric D, Darst RK, Fortunato S (2014) Community detection in networks: structural communities versus ground truth. Phys Rev E 90(6):062805
Hyland CC, Tao Y, Azizi L, Gerlach M, Peixoto TP, Altmann EG (2021) Multilayer networks for text analysis with multiple data types. EPJ Data Sci 10(1):33
Karrer B, Newman ME (2011) Stochastic blockmodels and community structure in networks. Phys Rev E 83(1):016107
Lancichinetti A, Sirer MI, Wang JX, Acuna D, Körding K, Amaral LAN (2015) A high-reproducibility and high-accuracy method for automated topic classification. Phys Rev X 5:011007
Mikolov T, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. Nips 1–9
Mikolov T, Corrado G, Chen K, Dean J (2013) Efficient estimation of word representations in vector space. In: Proceedings of the international conference on learning representations (iclr 2013), pp 1–12
Miller J (1981) Rg (1981): simultaneous statistical inference. Springer-Verlag
Paranyushkin D (2011) Identifying the pathways for meaning circulation using text network analysis. Nodus Labs 26
Paulheim H (2017) Knowledge graph refinement: a survey of approaches and evaluation methods. Semant Web 8(3):489–508
Peixoto TP (2015) Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. Phys Rev E 92(4):042807
Peixoto TP (2017) Nonparametric bayesian inference of the microcanonical stochastic block model. Phys Rev E 95:012317
Peixoto TP (2021) Revealing consensus and dissensus between network partitions. Phys Rev X 11(2):021003
Piantadosi ST (2014) Zipf's word frequency law in natural language: a critical review and future directions. Psychon Bull Rev 21:1112–1130
Rissanen J (1978) Modeling by shortest data description. Automatica 14(5):465–471
Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. Proceedings of the eighth ACM international conference on web search and data mining, pp 399–408
Simonetti A, Albano A, Plaia A, Tumminello M (2023) Ranking coherence in topic models using statistically validated networks. J Inf Sci 01655515221148369
Sprent P (2011) Fisher exact test. Lovric M (Eds) International encyclopedia of statistical science, pp 524–525. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2-253
Traag VA, Waltman L, Van Eck NJ (2019) From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9(1):1–12
Tumminello M, Consiglio A, Vassallo P, Cesari R, Farabullini F (2023) Insurance fraud detection: a statistically validated network approach. J Risk Insur 90(2):381–419
Tumminello M, Edling C, Liljeros F, Mantegna RN, Sarnecki J (2013) The phenomenology of specialization of criminal suspects. PLoS ONE 8(5):e64703
Tumminello M, Micciche S, Lillo F, Piilo J, Mantegna RN (2011) Statistically validated networks in bipartite complex systems. PLoS ONE 6(3):e17994
Tumminello M, Micciche S, Lillo F, Varho J, Piilo J, Mantegna RN (2011) Community characterization of heterogeneous complex systems. J Stat Mech Theory Exp 2011(01):P01019
Valles-Catala T, Massucci FA, Guimera R, Sales-Pardo M (2016) Multilayer stochastic block models reveal the multilayer structure of complex networks. Phys Rev X 6(1):011036
Veremyev A, Semenov A, Pasiliao EL, Boginski V (2019) Graph-based exploration and clustering analysis of semantic spaces. Appl Netw Sci 4:1–26
Xu Y, Lapata M (2019) Weakly supervised domain detection. Trans Assoc Comput Ling 7:581–596
Zhu Y, Yan X, Getoor L, Moore C (2013) Scalable text and link analysis with mixed-topic link models. In: Proceedings of the 19th ACM sigkdd international conference on knowledge discovery and data mining, pp 473–481
Zipf GK (1936) The psycho-biology of language: An introduction to dynamic philology, london: G. Routledge. INDEX BADIP 95:51–53
Zuo Y, Zhao J, Xu K (2016) Word network topic model: a simple but general solution for short and imbalanced texts. Knowl Inf Syst 48(2):379–398

Publisher's Note