




Multilayer networks for text analysis with multiple data types

Charles C. Hyland^{1*}, Yuanming Tao¹, Lamiae Azizi¹, Martin Gerlach², Tiago P. Peixoto^{3,4} and Eduardo G. Altmann^{1*} 

*Correspondence:
christopherhyland95@gmail.com;
eduardo.altmann@sydney.edu.au
¹School of Mathematics and
Statistics, The University of Sydney,
NSW, 2006, Sydney, Australia
Full list of author information is
available at the end of the article

Abstract

We are interested in the widespread problem of clustering documents and finding topics in large collections of written documents in the presence of metadata and hyperlinks. To tackle the challenge of accounting for these different types of datasets, we propose a novel framework based on Multilayer Networks and Stochastic Block Models. The main innovation of our approach over other techniques is that it applies the same non-parametric probabilistic framework to the different sources of datasets simultaneously. The key difference to other multilayer complex networks is the strong unbalance between the layers, with the average degree of different node types scaling differently with system size. We show that the latter observation is due to generic properties of text, such as Heaps' law, and strongly affects the inference of communities. We present and discuss the performance of our method in different datasets (hundreds of Wikipedia documents, thousands of scientific papers, and thousands of E-mails) showing that taking into account multiple types of information provides a more nuanced view on topic- and document-clusters and increases the ability to predict missing links.

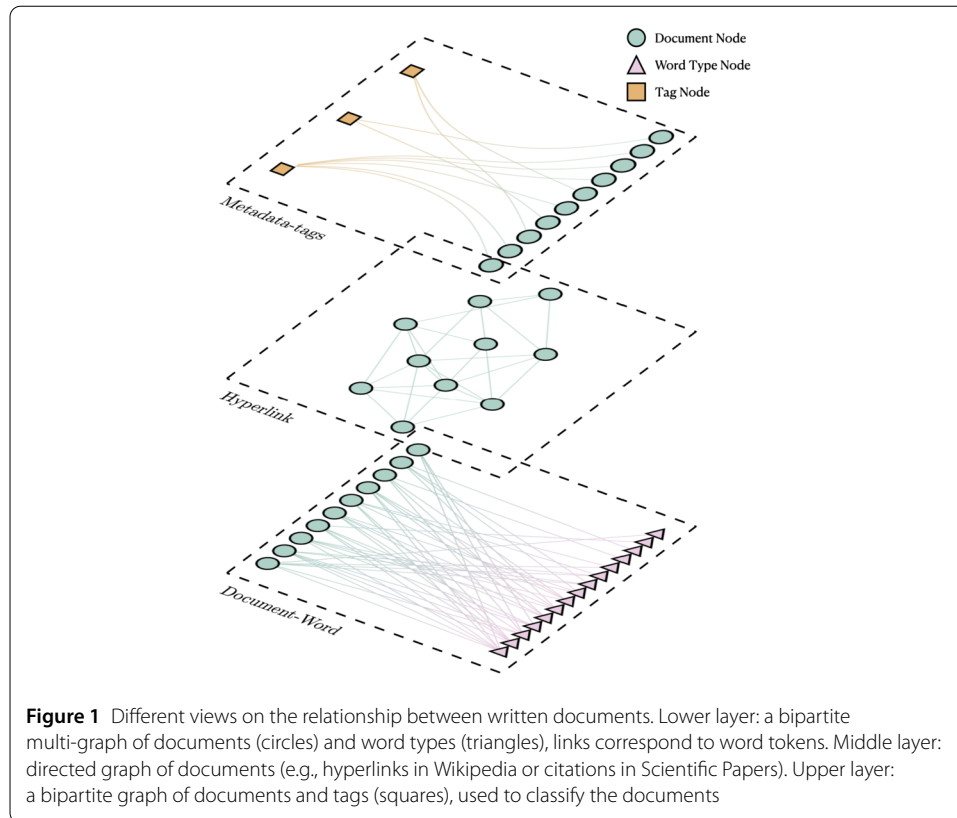
Keywords: Stochastic block models; Multilayer networks; Natural language processing; Complex systems; Data science

1 Introduction

A widespread problem in modern Data Science is how to combine multiple data types such as images, text, and numbers in a meaningful framework [1–5]. The traditional approach to tackle this challenge is to construct machine learning pipelines in which each data type is treated separately—sequentially or in parallel—and the partial results are combined at the end of the procedure [6, 7]. There are two problems with such a procedure. First, it leads to the development of ad-hoc solutions that are highly contingent on the dataset in question [8, 9]. Second, each model is trained independently from one another, meaning that the relationships between the different types of data are not taken into account [10, 11]. These problems show the need of developing a unified statistical framework applied simultaneously to the different types of data [12].

In this paper, we investigate the problem of clustering and finding topics in collections of written documents for which additional information is available as metadata and as hy-

© The Author(s) 2021. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.



perlinks between documents. We obtain a unified statistical framework to this problem by mapping it to the problem of inferring groups in multilayer networks. The key design for the unified framework proposed here is inspired by the connections [3, 13–15] between the problems of identifying (i) topics in a collection of written documents (i.e. topic modeling) [16] and (ii) communities in complex networks (i.e. community detection) [17]. In particular, Ref. [15] shows that both problems can be tackled using Stochastic Block Models (SBM) [3, 12, 18–22] and that SBMs, previously applied to find communities in complex networks, outperform and overcome many of the difficulties of the most popular unsupervised methods to infer structures from large collections of texts (topic modelling methods such as the Latent Dirichlet Allocation [23] and its generalizations). However, these approaches have been applied only to the textual part of collections of documents, ignoring additional information available about them. For instance, in datasets of scientific publications, one would consider only the text of the articles but not the citation network (used in traditional community detection methods [17]) or other metadata (such as the journal or bibliographical classification) [24, 25]. We propose here an extension of Ref. [15] and show how the diversity of information typically available about documents can be incorporated in the same framework by using multilayer SBMs [3, 10, 11]. As illustrated in Fig. 1, in addition to the bipartite Document-Word layer discussed in Ref. [15], here we incorporate a Hyperlink layer connecting the different written documents and a Metadata-Document layer that incorporates tags and other metadata. The key difference to other multilayer networks [4], as explored in Sect. 2 below, is that statistical laws [26] governing the frequency of words on documents leave fingerprints on the density of the different network layers. Our investigations in different datasets, reported in Sect. 3 for collection of Wikipedia

articles and in the Supplementary Information for three other datasets, reveal that the proposed multilayer approach leads to improved results when compared to both the topic modelling approach of [15] and the usual community detection of (hyperlink) networks. Our approach leads to a more nuanced view on the communities of documents, generates a list of topics associated to the communities, and improves the link-prediction capabilities when compared to the hyperlink network alone [27]. The details on our methods can be found in the appendices, Supplementary Information, and in the repository [28].

2 Multiple data sources as multilayer networks

In this section we introduce the general methodology of our paper: we introduce the types of data we are interested in (Sect. 2.1), we show how they can be represented as a multilayer network and discuss the properties of these networks (Sect. 2.2), and we describe how they can be modelled using Stochastic Block Models (Sect. 2.3).

2.1 Setting: multiple data sources

We consider a collection of $d = 1, \dots, D$ documents and we are interested in clustering and finding underlying similarities between them using combinations of the following information:

Text (T): Each document contains k_d word tokens from a vocabulary of V word types ($M = \sum_d k_d$ is the total number of word tokens).

Hyperlinks (H): Documents are linked to each other by building a (directed) graph or network.

Metadata (M): Documents are classified by tags or other metadata.

These characteristics are typical for textual data and networks. Here we explore three types of such datasets, summarized in Table 1. The main dataset we use to illustrate our results was extracted from the English Wikipedia, where the documents are articles (in scientific categories), the text is the content of the articles, hyperlinks are links between articles contained in the text, and metadata are tags introduced by users to classify the articles (categories). In our main example, we selected hundreds of articles in one of three scientific categories of Wikipedia (see Appendix 1 for details). Our main findings are confirmed in a second Wikipedia dataset (obtained choosing different scientific categories), in a citation dataset (documents are scientific papers, hyperlinks are citations, the text is extracted from the title and abstract, and metadata are scientific categories), and in an E-mail dataset (documents are all E-mails from the same user, hyperlinks correspond to E-mails sent between users, and the text is the content of the E-mails). These results and further details of the data are presented in the Supplementary Information 1 (see Additional file 1-Sect. 1).

2.2 Data as networks

The data described above can be represented as multilayer networks. The Hyperlink layer is the most obvious one, where documents are nodes and the hyperlinks are directed edges. The Metadata layer is built by a bipartite network consisting of metadata tags and documents as nodes, whereby undirected edges correspond to documents containing a given metadata-tag. Finally, the Text layer is obtained by restricting the text analysis to the level of word frequencies (bag-of-words) and then considering the bipartite network of word (types) and documents, where the edges correspond to word tokens (i.e., the count

Table 1 Summary of the datasets used in this paper

	Wikipedia Dataset in Manuscript	Wikipedia Dataset in SI	E-mail Dataset in SI	Citation dataset in SI
Nodes:				
Documents	120	316	4894	2542
Word Types	11,545	16,344	66,088	7677
Metadata Tags	Physics, Maths, Biology	Statistics, Maths, Electrical Engineering	0	52 Categories
Edges:				
Hyperlinks	309	1530	18,005	4590
Word Tokens	155,093	321,147	761,179	116,889
Tag Labels	120	316	0	2542

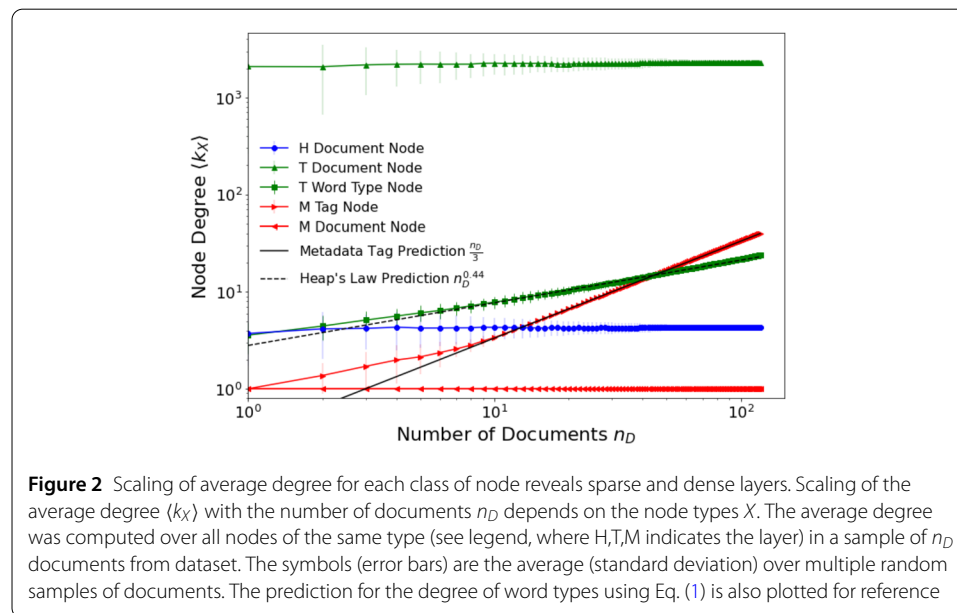


Figure 2 Scaling of average degree for each class of node reveals sparse and dense layers. Scaling of the average degree $\langle k_X \rangle$ with the number of documents n_D depends on the node types X . The average degree was computed over all nodes of the same type (see legend, where H,T,M indicates the layer) in a sample of n_D documents from dataset. The symbols (error bars) are the average (standard deviation) over multiple random samples of documents. The prediction for the degree of word types using Eq. (1) is also plotted for reference

of how often a word type appears in a document). While word-nodes and metadata tags appear only in the text and the metadata layer, all layers have document nodes in common. The novelty of our multilayer approach, in comparison to other approaches using multilayer networks, is the inclusion of the text layer. The importance of using a bipartite multigraph layer [22] to represent the text, instead of alternative “word networks” [14, 29, 30], is that it contains the complete information of word occurrence in documents and allows for a formal connection to topic-modelling methods [15, 20].

We now investigate the properties of the multilayer network described above, based on known results in networks and textual data. The most striking feature of this network is that the size of the different layers varies dramatically and scales differently with system size. A first indication of this lack of balance is seen by looking at the number of edges shown in Table 1: the number of edges in the text layer (i.e. word tokens) is substantially larger than the number of nodes or edges in all of the other layers. Such an imbalance is expected in all datasets in which the same type of data as outlined in Sect. 2.1 is present. To see this, we investigate in Fig. 2 how the average degree $\langle k_X \rangle$ (number of edges/ total number of nodes) of the different node types X scale with the number of documents n_D (which plays the role of system size). For the document nodes in the Hyperlink layer and the Text layer we see a constant average degree, typical of sparse networks. The Metadata

layer yields a trivial scaling linear with n_D as in dense networks because each document has one edge to a metadata node. More interestingly, the average degree of the word type nodes in the Text layer, $\langle k_V \rangle$, shows a growth that scales as

$$\langle k_V \rangle \sim n_D^\gamma, \quad (1)$$

with $0 < \gamma < 1$. This is between the usual limits expected for sparse ($\gamma = 0$) and dense ($\gamma = 1$) networks.

We now explain the observation in Eq. (1) in terms of properties of text in general. More specifically, the type-token relationship in texts follows Heaps' law [26, 31, 32], which states that the number of word types V scales with the word tokens M as

$$V \sim M^\beta, \quad (2)$$

whereby $0 < \beta < 1$ is the parameter of interest. The average degree is obtained as $\langle k_V \rangle = M/V$ and $n_D \propto M$ (where the proportionality constant is the average size of Wikipedia articles, in word tokens). Combining this with Eqs. (1) and (2) we obtain that $\gamma = 1 - \beta$. From the data used here, we estimate a Heaps' exponent $\beta = 0.56$, that leads to a prediction of $\gamma = 0.44$. This prediction is shown as a dashed line in Fig. 2 and is in good agreement (for large n_D) with the average degree of word nodes.

2.3 Stochastic block models

To achieve our goal of clustering documents and identifying topics considering multiple type of datasets simultaneously, we need to explore statistical patterns in the connectivity of the multilayer networks discussed above. This can be obtained using Stochastic Block Models (SBMs). The choice of SBMs is based on the existence of a successful computational and theoretical framework, reviewed in Ref. [12], that can be applied to networks with the characteristics needed in our problem: different types of networks (directed, bipartite, and multi edges), multilayer networks [11], and accounting for key ingredients to detect communities (e.g., degree correction and a nested/hierarchical generalizations [33]). Our previous analysis of bipartite word-document networks using this framework has outperformed traditional topic modelling approaches [15].

SBMs are a family of random-graph models that generate networks with adjacency matrix A_{ij} with probability $P(\mathbf{A}|\mathbf{b})$, where the vector \mathbf{b} with entries $b_i \in \{1, \dots, B\}$, specifies the membership of nodes $i = 1, \dots, D$ into one of B possible groups. For our multilayer network design—developed for the three types of data (H,T,M) as discussed in Sect. 2.2—we fit the SBM framework to each layer combining them by constraining document groups to be the same across all layers, i.e. with a joint probability

$$P(\mathbf{A}_H, \mathbf{A}_T, \mathbf{A}_M|\mathbf{b}) = P(\mathbf{A}_H|\mathbf{b})P(\mathbf{A}_T|\mathbf{b})P(\mathbf{A}_M|\mathbf{b}), \quad (3)$$

where \mathbf{A}_H , \mathbf{A}_T and \mathbf{A}_M are the adjacency matrices of each respective layer. In each individual layer, edges between nodes i and j are sampled from a Poisson distribution with average [20]

$$\theta_i \theta_j \omega_{b_i, b_j}, \quad (4)$$

whereby ω_{rs} is the expected number of edges between group r and s , b_i is the group membership of node i , and θ_i is overall propensity with which a node is selected within its own group. Non-informative priors are used for the parameters θ and ω and the marginal likelihood of the SBM is computed as [34]

$$P(\mathbf{A}|\mathbf{b}) = \int P(\mathbf{A}|\omega, \theta, \mathbf{b})P(\omega, \theta|\mathbf{b}) d\theta d\omega, \quad (5)$$

Based on this, we consider the overall posterior distribution for a single partition conditioned the edges on all layers [35]

$$P(\mathbf{b}|\mathbf{A}_H, \mathbf{A}_T, \mathbf{A}_M) = \frac{P(\mathbf{A}_H|\mathbf{b})P(\mathbf{A}_T|\mathbf{b})P(\mathbf{A}_M|\mathbf{b})P(\mathbf{b})}{P(\mathbf{A}_H, \mathbf{A}_T, \mathbf{A}_M)}. \quad (6)$$

With this approach, not only the words but also the documents are now clustered into categories. We implement the inference using the package `graph-tool` [28, 36–38] (see Additional file 1-Sect. 2 for details and Ref. [28] for our codes).

3 Application to Wikipedia data

In this section we apply the methodology and ideas discussed above to the Wikipedia dataset which contains articles classified by users in the categories Mathematics, Physics, and Biology. We are interested in comparing the outcomes and performance of the models discussed above applied to the different types of information in the data. We fit multiple variants of the multilayer SBM, whereby we choose different layers to be included in the model.

3.1 Description length

The performance of each model can be measured by the extent to which a model succeeds in describing (compressing) the data. This can be quantified computing its description length (DL) [39, 40]

$$\text{DL} = -\log P(\mathbf{A}_H, \mathbf{A}_T, \mathbf{A}_M, \mathbf{b}), \quad (7)$$

which describes the information necessary to describe both the data and the model parameters. From Eq. (6), we see that minimizing the description length is equivalent to maximizing the posterior probability $P(\mathbf{b}|\mathbf{A}_H, \mathbf{A}_T, \mathbf{A}_M)$.

In Table 2 we summarise the DL obtained for each model in our dataset. It is quite clear that the DL of the models containing the Text layer are much larger than those containing only the Hyperlink and Metadata layers. This is a direct consequence of the large number of word types in the data, when compared to documents or hyperlinks, the lack of balance between the layers mentioned in Sect. 2.2. This lack of balance between layers thus has important consequences for the inference of partitions and our ability to compare the different models. For instance, the effectiveness of the multilayer approach could be evaluated by comparing the DL of the multilayer model (e.g., DL of model $H + T$) to the sum of the DL of the single-layer models (e.g., DL of model H + DL of model T). In our case this comparison is not very informative because the DL of the combined model is dominated by the largest layer and the DL of the small layer often lies within the fluctuations obtained from multiple MCMC runs (see Additional file 1-Sect. 2). This reasoning suggests

Table 2 Description length for each combination of layers in the multilayer stochastic block model. We compute the average description length (DL), Eq. (7), for each model class alongside the standard deviation over multiple MCMC runs. We also retrieved the minimum DL (MDL) over all the runs. The DL of the Text layer exceeds the Hyperlink and Metadata layer by several orders of magnitude, thus contributing the most to the Hyperlink + Text model

Model	Layers	DL	MDL
H	Hyperlink	1135 (0)	1135
T	Text	257,775 (471)	256,973
M	Metadata	76 (0)	76
H + M	Hyperlink + Metadata	1295 (18)	1281
H + T	Hyperlink + Text	270,230 (2228)	267,102
H + T + M	Hyperlink + Text + Metadata	282,560 (624)	281,133

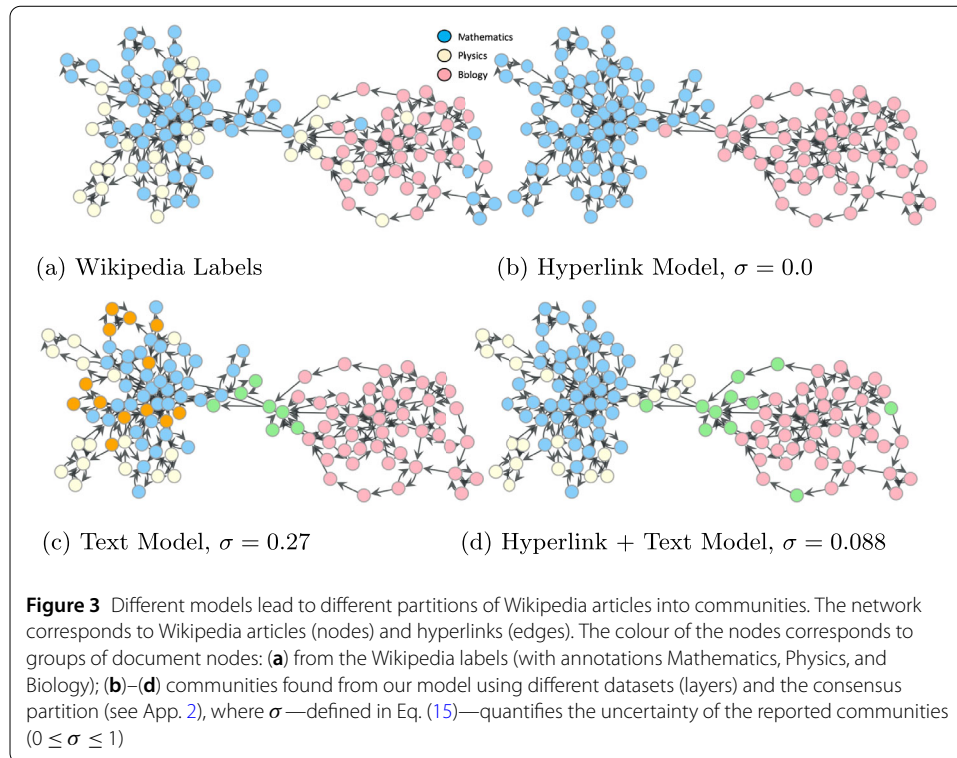
that the clustering of nodes would be dominated by the Text layer or, if the Text layer is excluded, by the Hyperlink layer which will dominate over the Metadata layer.¹ However, we will see below that there are still significant and meaningful differences in the clustering of nodes obtained using different combinations of layers. This happens because the inference problem remains non-trivial because the DL landscape contains many distinct states with similar values in the DL so that even small effects due to the H and M layers can affect the outcome.

3.2 Qualitative comparison of groups of documents

Community detection methods aim to find the partition of the nodes that best captures the structure in the network in a meaningful way whilst being robust to noise [12, 21]. We thus evaluate the different models by comparing the resulting partitioning of documents [35]. Specifically, we fit the Hyperlink, Text, and Hyperlink + Text model and obtain a best partition from combining multiple samples from the posterior $P(\mathbf{b}|\mathbf{A})$ for each model to construct a point estimate, which utilises the different parts of the posterior distribution. We then project the group membership onto the Hyperlink layer (which only contains document-nodes) and retrieve the consensus partition alongside the uncertainty of the partition [41] (see Appendix 2 for details).

Our results are shown in Fig. 3 and reveal that our model is successful in retrieving different meaningful groupings of the articles depending on the available data (i.e. layers included in the model). We first notice that the classification of articles made by users—panel (a), Wikipedia label—group articles in Mathematics and Biology that are strongly linked with each other (through hyperlinks), whereas Physics articles appear intertwined in between them. When we infer the partition of nodes based only on the hyperlink network—panel (b), Hyperlink model—we obtain that our model obtains 2 groups and it is quite confident about it (uncertainty is zero, $\sigma = 0.$). This partition resembles the partition based on Wikipedia labels. When the documents are partitioned based on their text—panel (c), Text Model-, a richer picture emerges. There is a large community that resembles closely the documents classified as Biology and one of the communities obtained using the hyperlinks layer. However, the remaining documents (most of the Mathematics and Physics articles) are now grouped in 4 categories (i.e., 5 communities in total) which are still linked

¹In the Metadata layer, we found that the metadata tags will form a trivial single group as there is insufficient evidence for the model to construct more than one group. Therefore, we constrained metadata tags to be in separate groups to ensure that they provide additional information to the models being fitted.

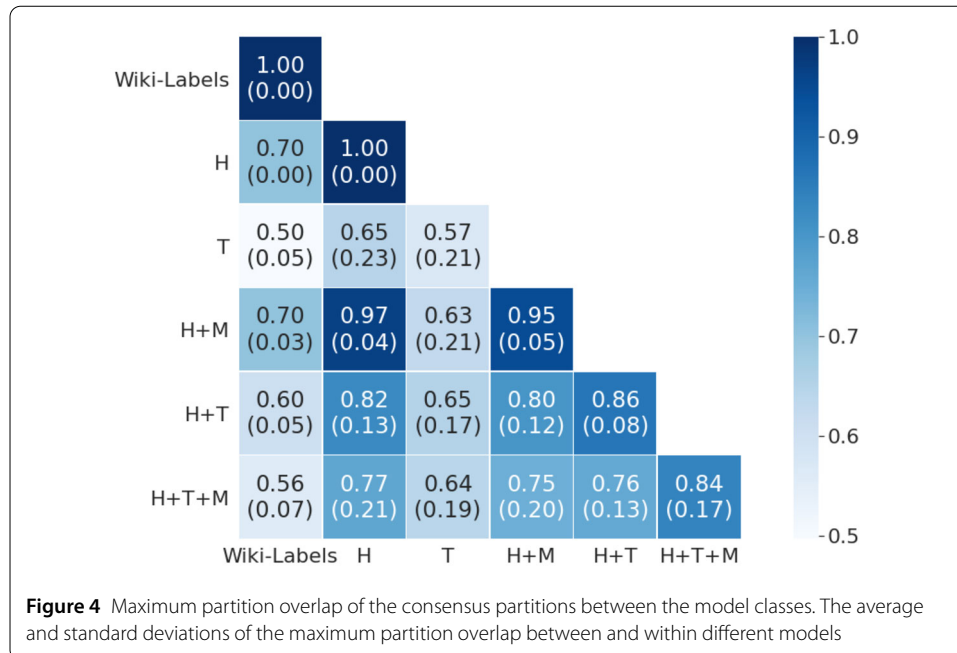


to each other but more loosely than before (even though Fig. 3 shows the Hyperlink network, the Hyperlinks were not used to group documents in panels a) and c)). Finally, when hyperlinks and text are used simultaneously—panel (d)—4 communities are found, which resemble the previous ones but that also show important distinctions. This demonstrates that even if the Text layer dominates the description length, there are noticeable differences in the inferred partitions when using the hyperlinks in addition to text for clustering documents.

We now argue that the more nuanced classification of documents obtained with the Text and Hyperlink + Text models are qualitatively meaningful. For example, we can see a cluster of 5 (Physics) nodes in the bottom left of the Hyperlink model that was not identified as a separate group, but it is now picked up in the Text and Hyperlink + Text model. This cluster of nodes include Wikipedia articles on the Josephson effect, macroscopic quantum phenomena, magnetic flux quantum, macroscopic quantum self trapping, and quantum tunnelling. Even more strikingly, in the bottom of the network there is a lone (Physics) green node surrounded by (Biology) red nodes which corresponds to the Wikipedia article on isotopic labelling (a technique in the intersection of Physics and Biology). In traditional community detection methods, which use link information as an indicator of groups, such a node would be in the community of its surrounding neighbours. However, in the Hyperlink + Text model, we are able to detect the uniqueness of such a node.

3.3 Quantitative comparison between different models

In the example discussed above it was clear that the different models yielded different yet related partitions of Wikipedia articles. In order to quantify the similarity of the results of the different models, we performed a systematic comparison of the partitions generated by multiple runs of each model and computed their similarities using the maximum



overlap partition (Fig. 4, see Appendix 2 for details). The results show that the partitions generated by the Hyperlink + Text model is most similar to the Text model. Similar results are obtained in our alternative datasets—see Additional file 1-Sect. 1—and using the normalised mutual information (NMI) as an alternative dissimilarity measure—see Additional file 1-Sect. 3.

We also compare the Hyperlink and Hyperlink + Text model in terms of their ability to predict missing edges [27, 42] (see Appendix 3 for details on our method). We found that the Hyperlink + Text model has an Area-Under-Curve (AUC) score of 0.63 ± 0.06 (average \pm standard deviation) and the Hyperlink model has 0.54 ± 0.02 , with the difference being statistically significant ($p = 0.0013$, using a 2-sample t-test). This confirms that the multilayer approach proposed here is successful in retrieving existing relationships that are missed in the network-only approach.

3.4 Lack of balance in the hyperlink-text model

The results of the previous sections are strongly influenced by the lack of balance in Hyperlink + Text model, as discussed in Sect. 2. To further illustrate this point, here we artificially reduce the unbalance of the multilayer network by sampling a fraction μ of word tokens before fitting a Hyperlink + Text model. We expect that, as we increase the fraction of words μ , the Text layer will increasingly dominate the inference. This expectation is confirmed in Fig. 5, which shows that for $\mu \geq 0.6$, the partition overlap of the μ -hyperlink-text model is statistically indistinguishable from the partition overlap obtained using the Text-only model. That is, we see that the Text layer dominates the inference in the μ -Hyperlink + Text for $\mu \geq 0.6$. However, as discussed above, the effect of the hyperlink layer can lead to different consensus partition.

3.5 Topic modelling: groups of words

Since our approach provides a clustering of all nodes, we not only group documents but also words. The groups of word (types) can be interpreted as the topics of the documents

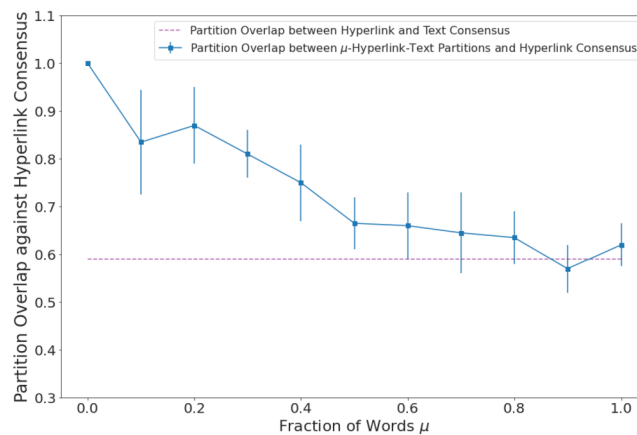


Figure 5 Text layer determines the partitions obtained in the multilayer model (Hyperlink + Text). Similarity (overlap of consensus partition) between Hyperlink partition and μ -Hyperlink-Text partition as a function of the subsampling parameter μ where $\mu = 0$ ($\mu = 1$) corresponds to the case with all (none) of the word tokens removed in the Hyperlink-Text model. For a given value of μ , a random fraction $1 - \mu$ of the words were removed and the Hyperlink + Text model was then fitted for multiple iterations. The consensus partition was then computed for the Hyperlink + Text model and its partition overlap with Hyperlink model. A higher sub-sampling of text (i.e. smaller values of μ) results in the consensus partition between the Hyperlink + Text and Hyperlink model having a high degree of overlap

linked to them, showing that our framework simultaneously solves the traditional problem of topic modeling [14, 23]. Below we show the topics obtained in our Wikipedia dataset, as an example of our generic topic-modelling methodology.

In the consensus partition of the Hyperlink-Text network (see Fig. 3) we found 12 topics (groups of word types). The most frequent words in each of these topics is shown in Table 3. Qualitatively, we see that topics are often composed of semantically related words, e.g. topics 1 and 3 contain a large number of key words associated to Biology whilst topics 5 and 10 contains a large number of jargon related to Physics.

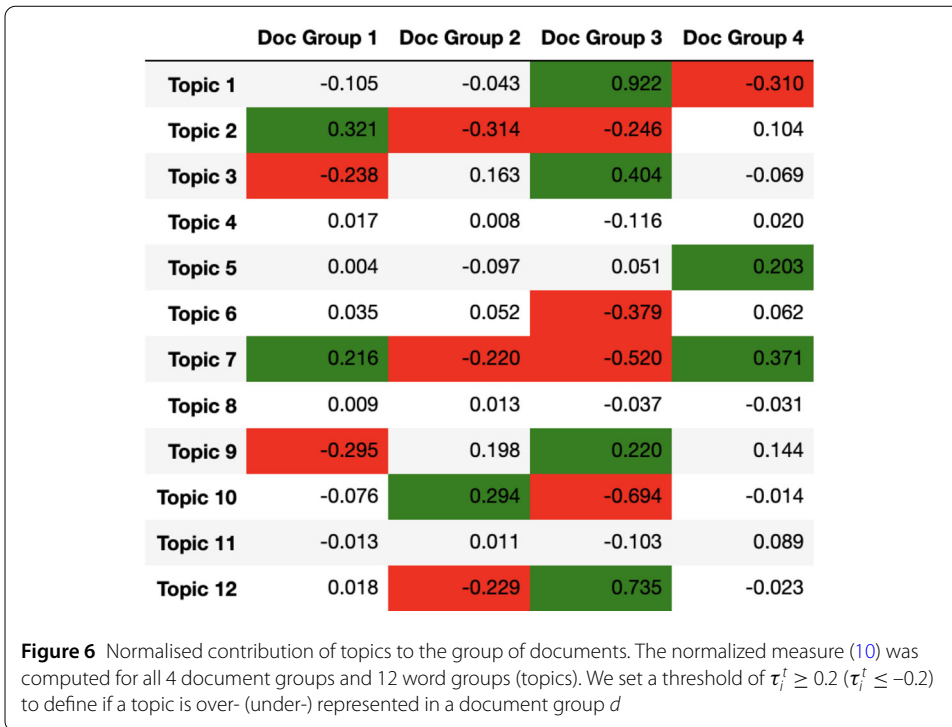
We now discuss the topical composition of (groups of) documents. Let $T = B_V$ be the number of topics and B_D be the number of document groups, then the mixture proportion of topic $t = 1, \dots, T$ in document group $i = 1, \dots, B_D$ is given by

$$f_i^t = \frac{n_i^t}{\sum_{t'=1}^T n_i^{t'}}, \quad (8)$$

where n_i^t is the number of word tokens in topic t that appeared in documents d in document-group i . The results obtained for the four document groups are shown at the bottom of Table 3. Interestingly, topic 4 cannot be identified with any specific group of documents. This suggests that the words in this topic are similar to so-called stopwords, a pre-defined set of common words considered uninformative which are typically removed from the corpus before any model is to be fitted in order to improve the model [43]. This is consistent with the finding of Ref. [15] that SBMs applied to word-document networks were able to automatically filter stop words by grouping them into a “topic” that is well connected to all documents. Our findings suggest that the same is true for multilayer models and that our approach is robust against the presence of stopwords. In fact, this stopword topic is responsible for a large fraction (40%) of the topic-proportion for all groups of

Table 3 Groups of word types as topics. Upper table: the 20 most frequent words in the 12 topics (word groups) found in the consensus partition of the Hyperlink + Text model. Lower table: the topic proportion (8) of the four groups of documents

Group	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
	occur specific within process produce increase mechanism cause acid target step bind gene control human encode body dna rate formation	chemical loop electron transition chemistry atom stable ion pattern crystal reach prediction label compound reaction crystallography oxygen solid unstable atomic	protein cell site activity pathway cellular organism amino enzyme synthesis species tissue proteins cancer activate genetic membrane release mutation regulation	one also use form function give two may however example result call first different know make number structure show include	system time state work second effect phase interaction potential could derive difference full free hold observe research propose year paper	group space define field point theory value product constant vector physical physic energy element particle parameter whose mathematical linear matrix	transform law critical magnetic matter scale plane spectrum volume temperature axis statistical sign effective symbol speed electric equilibrium unchanged assumption	action translation class line measure central primary unknown interior surface side center flow copy block configuration read alternate table transport	happen frequency fast output input algorithm digital	quantum equation transformation theorem symmetry dimension denote classical let mechanic coordinate lie real mathematics geometry operator differential representation invariant generalize	every basis unique degree positive open close fix reference exact closed restrict picture infinite index interpretation formulate equivalently maximal scheme	factor expression region molecular sequence information end molecule domain signal alternative bond family interact cycle biological life translate structural biology
1 (Mathematics)	0.1481	0.0213	0.0444	0.400	0.0578	0.0722	0.0510	0.0127	0.00163	0.147	0.0105	0.0338
2 (Physics)	0.138	0.0409	0.0291	0.404	0.0643	0.0710	0.07940	0.0127	0.0009583	0.105	0.0102	0.0445
3 (Biology)	0.297	0.0234	0.0536	0.351	0.0673	0.0426	0.0313	0.0121	0.00166	0.0347	0.00929	0.0759
4 (New Group)	0.107	0.0342	0.0355	0.405	0.0771	0.0728	0.0895	0.0122	0.00156	0.112	0.0113	0.0428



documents. The underlying reason for this is the higher frequency of these words, which (due to Zipf’s law) dominate the weights of the topic mixture models [44]. To overcome this feature, and assess the over- or under-representation of topics more rigorously, we account for the overall frequency of occurrence of words in topics t as

$$\langle f^t \rangle = \frac{\sum_{i=1}^{B_D} n_i^t}{\sum_{t=1}^T \sum_{j=1}^{B_D} n_j^t}, \tag{9}$$

and define the normalised value of the mixture proportion of topic t in document group i as

$$\tau_i^t = \frac{f_i^t - \langle f^t \rangle}{\langle f^t \rangle}. \tag{10}$$

This normalised measure has an intuitive interpretation: $\tau_i^t > 0$ ($\tau_i^t < 0$) implies that topic t is over-represented (under-represented) in document group d . In Fig. 6, we show τ_i^t for the 12 topics and the 4 document groups, providing a much clearer view on the connection between topics and groups of documents. For example, we see that document group 2 (articles labelled as Physics) has a large over-representation of topic 10, which corresponds to the Physics topic whilst being underrepresented in document group 2 (articles labelled as Biology). Looking at the model’s newly proposed document group (group 4) we see that it has an over-representation from topics 7 and 5 (and in a less extent from topics 2, 9, and 11), confirming its hybrid category.

4 Discussion and conclusions

In this paper, we introduced and explored a formal methodology that combines multiple data types (e.g., text, metadata, links) to perform the common tasks of clustering and infer-

ring latent relationships between documents in text analysis. The main theoretical advantage of our methodology is that it incorporates all the different types of data into a single, consistent, statistical model. Our approach is based on an extension of multilayer Stochastic Block Models, that have been used previously to find communities in (sparse) complex networks and that is used here to perform text analysis (see Refs. [3, 18] for alternative uses of SBMs for topic modelling). On the one hand, our method extends community-detection methods to the analysis of text in the presence of multiple data types, our main finding being that: (i) universal statistical properties of texts lead to different link densities at the different layers of the network; and (ii) that the word layer plays a dominant role in the inference of partitions. On the other hand, our method can be viewed as a generalized topic modelling method that incorporates meta-data and hyperlinks, labels the communities of documents by examining the proportion of topics, and builds on the previous connections between SBMs and Latent Dirichlet Allocation [15, 20].

Our investigations on four different datasets show consistent results that reveal the potential and limitations of our approach. Our most important finding is that our methodology succeeds in using the multiple data types (e.g., a text layer) leading to more nuanced communities of documents and in increasing the ability to predict missing links. On the practical side, the lack of balance between the different layers poses challenges on how to evaluate the contributions of different layers because the description length obtained in the inference process is dominated by the text layer and variations obtained within the (Monte Carlo) inference process become larger than the contribution of alternative layers. This suggests further investigations on the role of unbalanced layers in multilayer networks, and how to deal with them within the proposed framework, as important steps to expand the success of complex-network methods to other classes of relevant datasets.

Appendix 1: Wikipedia data collection and preparation

We used a snapshot of the Wikipedia data retrieved on the 5th of June, 2020. The following lists the data extraction and processing steps:

1. *Data Retrieval*: We retrieved the Wikipedia articles and their content (metadata, text, link) through the MediaWiki API² and parsing the Wikipedia dumps³.
2. *Network Formulation*: We constructed a network whereby each node represents a Wikipedia article and each (directed) edge represents hyperlinks between the Wikipedia articles. We removed any nodes with less than 2 outgoing links.
3. *Retrieve Connected Component*: For ease of analysis, we extracted the largest connected component in the hyperlink network constructed.
4. *Text Processing*: We process the Wikipedia text data through both tokenization and lemmatization using NLTK [45].

The resultant dataset is available in our repository [28].

²<https://en.wikipedia.org/w/api.php>

³<https://dumps.wikimedia.org/>

Appendix 2: Maximum overlap and consensus partition

The maximum overlap between partitions measures the similarities between sets of partitions. The maximum overlap between partitions \mathbf{x} and \mathbf{y} is given by

$$w(\mathbf{x}, \mathbf{y}) = \arg \max_{\mu} \sum_i \delta_{x_i, \mu(y_i)}, \quad (11)$$

where μ is a bijective mapping between the group labels [41].

The normalized maximum overlap between partitions \mathbf{x} and \mathbf{y} is given by

$$w(\mathbf{x}, \mathbf{y}) = \frac{\sum_i \delta_{x_i, y_i}}{N}, \quad (12)$$

where N is the number of nodes and lies in the unit interval $[0,1]$.

Given multiple partitions, we also wish to extract a consensus partition $\hat{\mathbf{b}}$ which has the maximal sum of overlaps with all the partitions. Such a consensus partition can be obtained through the double maximization of the set of equations:

$$\hat{b}_i = \arg \max_r \sum_m \delta_{\mu_m(b_i^m), r}, \quad (13)$$

$$\mu_m = \arg \max_{\mu} \sum_r m_{r, \mu(r)}^{(m)}, \quad (14)$$

where μ is a bijective mapping between the group labels and $m_{r, \mu(r)}^{(m)}$ is the contingency table between $\hat{\mathbf{b}}$ and partition $\mathbf{b}^{(m)}$. An iterative procedure is then carried out on the set of equations until no further improvement is possible. The uncertainty σ of the consensus partition obtained from M_p partitions is quantified as [28, 41]

$$\sigma = 1 - \frac{1}{NM_p} \sum_i \sum_m \delta_{\mu_m(b_i^m), \hat{b}_i}. \quad (15)$$

Appendix 3: Supervised learning via link prediction

A supervised learning approach to select the best model can be done through the task of link prediction [42, 46]. Let \mathbf{A}^O be the observed network and $\delta\mathbf{A}$ be missing or spurious edges. The desired posterior distribution of missing entries $\delta\mathbf{A}$ conditioned on the observed network \mathbf{A}^O can be computed as

$$P(\delta\mathbf{A}|\mathbf{A}^O) = \frac{\sum_{\mathbf{b}} P(\mathbf{A}^O \cup \delta\mathbf{A}|\mathbf{b})P(\mathbf{b}|\mathbf{A}^O)}{P(\mathbf{A}^O|\mathbf{b})}. \quad (16)$$

However, as the normalization constant is difficult to obtain, the numerator of Eq. (16) can be computed by sampling partitions from the posterior and then inserting or deleting edges from the graph and computing the new likelihood. As a result, we therefore may compute the relative probability between specific sets of alternative predictive hypotheses $\{\delta\mathbf{A}_i\}$ through the likelihood ratios ratio

$$\lambda_i = \frac{P(\delta\mathbf{A}_i|\mathbf{A}^O)}{\sum_j P(\delta\mathbf{A}_j|\mathbf{A}^O)}. \quad (17)$$

We can compute the area under curve (AUC) of the receiver operating characteristic curve to evaluate the SBM's classification abilities. Furthermore, given two sets of AUCs from two different models, we can compare the models' performance by computing the t-statistic for a null model with zero mean for the difference in AUC which is given by

$$t_{\Delta\text{AUC}} = \frac{\langle \Delta\text{AUC} \rangle}{\sigma_{\Delta\text{AUC}} / \sqrt{n}}, \quad (18)$$

where $\langle \Delta\text{AUC} \rangle$, $\sigma_{\Delta\text{AUC}}$, and n are the mean, standard deviation and size of the population.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1140/epjds/s13688-021-00288-5>.

Additional file 1. Supplementary information (PDF 600 kB)

Acknowledgements

Funding from The University of Sydney was received through the CTDS incubator scheme.

Availability of data and materials

Codes are available at <https://topsbm.github.io> and the datasets at https://github.com/martingerlach/hSBM_Topicmodel/tree/master/data.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MG, TP, and EGA conceived the idea. CCH and YT performed the numerical investigations, data analysis, and prepared the figures. All authors contributed to the methodological development. CCH, LA, MG, TPP, and EGA wrote the manuscript. All authors read and approved the final manuscript.

Author details

¹School of Mathematics and Statistics, The University of Sydney, NSW, 2006, Sydney, Australia. ²Wikimedia Foundation, San Francisco, USA. ³Department of Network and Data Science, Central European University, Quellenstraße 51, 1100, Vienna, Austria. ⁴Department of Mathematical Sciences, University of Bath, Claverton Down, BA2 7AY, Bath, United Kingdom.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 5 March 2021 Accepted: 14 June 2021 Published online: 28 June 2021

References

1. Kedem B, De Oliveira V, Sverchkov M (2017) Statistical data fusion. World Scientific, Singapore
2. Costanedo F (2013) A review of data fusion techniques. *Sci World J* 2013:704504
3. 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
4. Kivelä M, Arenas A, Barthélemy M, Gleeson J, Moreno Y, Porter M (2014) Multilayer networks. *J Complex Netw* 2(3):203–271
5. Zanin M, Papo D, Sousa PA, Menasalvas E, Nicchi A, Kubik E, Boccaletti S (2016) Combining complex networks and data mining: why and how. *Phys Rep* 635:1–44
6. Breck E, Zinkevich M, Polyzotis N, Whang S, Roy S (2019) Data validation for machine learning. In: Proceedings of SysML
7. O'Leary K, Uchida M (2020) Common problems with creating machine learning pipelines from existing code. In: Third conference on machine learning and systems (MLSys)
8. Arun R, Suresh V, Madhavan CEV, Murthy MNN (2010) On finding the natural number of topics with latent Dirichlet allocation: some observations. In: Advances in knowledge discovery and data mining, 391–402
9. Cao J, Xia T, Li J, Zhang Y, Tang S (2009) A density-based method for adaptive LDA model selection. *Neurocomputing* 72:1775–1781
10. Vallès-Català T, Massucci FA, Guimerà R, Sales-Pardo M (2016) Multilayer stochastic block models reveal the multilayer structure of complex networks. *Phys Rev X* 6:011036
11. Peixoto TP (2015) Inferring the mesoscale structure of layered, edge-valued and time-varying networks. *Phys Rev E* 92(4):042807
12. Peixoto TP (2019) Bayesian stochastic blockmodeling. In: Advances in network clustering and blockmodeling, ch. 11

13. Ball B, Karrer B, Newman MEJ (2011) Efficient and principled method for detecting communities in networks. *Phys Rev E* 84:036103
14. Lancichinetti A, Sirer MI, Wang JX, Acuna D, Körding K, Amaral LAN (2015) High-reproducibility and high-accuracy method for automated topic classification. *Phys Rev X* 5(1):011007
15. Gerlach M, Peixoto TP, Altmann EG (2018) A network approach to topic models. *Sci Adv* 4:eaaq1360
16. Blei DM (2012) Probabilistic topic models. *Commun ACM* 55
17. Fortunato S (2010) Community detection in graphs. *Phys Rep* 486(3–5):75–174
18. Bouveyron C, Latouche P, Zreik R (2016) The stochastic topic block model for the clustering of vertices in networks with textual edges. *Stat Comput*: 1–21
19. Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: first steps. *Soc Netw* 5(2)
20. Karrer B, Newman MEJ (2011) Stochastic blockmodels and community structure in networks. *Phys Rev E* 83:016107
21. Hastings M (2006) Community detection as an inference problem, physical review. *Phys Rev E, Stat Nonlinear Soft Matter Phys* 74:035102
22. Yen T-C, Larremore DB (2020) Community detection in bipartite networks with stochastic blockmodels. *Phys Rev E* 102:032309
23. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J Mach Learn Res* 3
24. Hric D, Peixoto TP, Fortunato S (2016) Network structure, metadata, and the prediction of missing nodes and annotations. *Phys Rev X* 6(3):031038
25. Newman M, Clauset A (2015) Structure and inference in annotated networks. *Nat Commun* 7
26. Altmann EG, Gerlach M (2016) Statistical laws in linguistics. *Creativity and universality in language*: 7–26
27. Guimerà R, Pardo MS (2009) Missing and spurious interactions and the reconstruction of complex networks. *Proc Natl Acad Sci* 106:22073–22078
28. Codes: TopSBM (Topic Models based on Stochastic Block Models, <https://topsbm.github.io>) and graph-tool (Efficient network analysis, <https://graph-tool.skewed.de>)
29. de Arruda HF, Costa LDF, Amancio DR (2016) Topic segmentation via community detection in complex networks. *Chaos* 26(6):063120
30. Leydesdorff L, Nerghes A (2017) Co-word maps and topic modeling: a comparison using small and medium-sized corpora ($N < 1000$). *Journal of the Association for Information Science and Technology* 68(4)
31. Herdan G (1960) Type-token mathematics. Mouton
32. Heaps HS (1978) *Information retrieval*. Academic, New York
33. Peixoto TP (2014) Hierarchical block structures and high-resolution model selection in large networks. *Phys Rev X* 4(1):011047
34. Peixoto TP (2017) Nonparametric Bayesian inference of the microcanonical stochastic block model. *Phys Rev E* 95(1):012317
35. Hric D, Darst RK, Fortunato S (2014) Community detection in networks: structural communities versus ground truth. *Phys Rev E* 90:062805
36. Peixoto TP (2014) Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. *Phys Rev E* 89(1):012804
37. Peixoto TP (2020) Merge-split Markov chain Monte Carlo for community detection. *Phys Rev E* 102:012305
38. Newman MEJ, Barkema GT (1999) *Monte Carlo methods in statistical physics*. Oxford University Press, London
39. Rissanen J (1978) Modeling by shortest data description. *Automatica* 14:465–471
40. Grünwald P (2007) *The minimum description length principle*. MIT Press, Cambridge
41. Peixoto TP (2021) Revealing consensus and dissensus between network partitions. *Phys Rev X* 11:021003
42. Vallès-Català T, Peixoto TP, Guimerà R, Sales-Pardo M (2018) Consistencies and inconsistencies between model selection and link prediction in networks. *Phys Rev E* 97:062316
43. Haddi E, Liu X, Shi Y (2013) The role of text pre-processing in sentiment analysis. *Procedia computer science*, vol 17
44. Altmann EG, Dias L, Gerlach M (2017) Generalized entropies and the similarity of texts. *J Stat Mech Theory Exp* 2017(1):014002
45. Bird S, Loper E, Klein E (2009) *Natural language processing with Python*. O'Reilly Media Inc.
46. Clauset A, Moore C, Newman MEJ (2008) Hierarchical structure and the prediction of missing links in networks. *Nature* 453:98–101

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)
