



Topic segmentation via community detection in complex networks

Henrique F. de Arruda, Luciano da F. Costa, and Diego R. Amancio

Citation: *Chaos* **26**, 063120 (2016); doi: 10.1063/1.4954215

View online: <http://dx.doi.org/10.1063/1.4954215>

View Table of Contents: <http://scitation.aip.org/content/aip/journal/chaos/26/6?ver=pdfcov>

Published by the [AIP Publishing](#)

Articles you may be interested in

[Simulating synchronization in neuronal networks](#)

Am. J. Phys. **84**, 467 (2016); 10.1119/1.4945009

[Generative model selection using a scalable and size-independent complex network classifier](#)

Chaos **23**, 043127 (2013); 10.1063/1.4840235

[Stability analysis and synchronization in discrete-time complex networks with delayed coupling](#)

Chaos **23**, 043108 (2013); 10.1063/1.4825095

[The Irony of Craigslist](#)

Comput. Sci. Eng. **10**, 88 (2008); 10.1109/MCSE.2008.64

[Web resources for teaching introductory modern physics: Two resources for teaching introductory atomic, wave and quantum physics](#)

Phys. Teach. **43**, 316 (2005); 10.1119/1.1903830



Topic segmentation via community detection in complex networks

Henrique F. de Arruda,¹ Luciano da F. Costa,² and Diego R. Amancio^{1,a)}

¹*Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, São Paulo, Brazil*

²*São Carlos Institute of Physics, University of São Paulo, São Carlos, São Paulo, Brazil*

(Received 8 December 2015; accepted 7 June 2016; published online 24 June 2016)

Many real systems have been modeled in terms of network concepts, and written texts are a particular example of information networks. In recent years, the use of network methods to analyze language has allowed the discovery of several interesting effects, including the proposition of novel models to explain the emergence of fundamental universal patterns. While syntactical networks, one of the most prevalent networked models of written texts, display both scale-free and small-world properties, such a representation fails in capturing other textual features, such as the organization in topics or subjects. We propose a novel network representation whose main purpose is to capture the semantical relationships of words in a simple way. To do so, we link all words co-occurring in the same semantic context, which is defined in a threefold way. We show that the proposed representations favor the emergence of communities of semantically related words, and this feature may be used to identify relevant topics. The proposed methodology to detect topics was applied to segment selected Wikipedia articles. We found that, in general, our methods outperform traditional bag-of-words representations, which suggests that a high-level textual representation may be useful to study the semantical features of texts. *Published by AIP Publishing.*

[<http://dx.doi.org/10.1063/1.4954215>]

Complex networks have been used to model and characterize many real systems, including written texts and linguistic phenomena. In this context, we have proposed novel network representations to tackle the topic segmentation task, an important text classification problem. Unlike traditional methods, we have found that written texts display a modular structure, as a consequence of their semantical organization in topics and subtopics. To our knowledge, this is the first time that such structure has been found in networks of words interaction. Using an optimized representation of the proposed model, we devised a strategy to cluster paragraphs on the same topic. Our results show that the use of connectivity information favors the identification of topics in texts, as in several cases the proposed models outperform traditional bag-of-word (BOW) representations. Our findings pave the way for the application in other related linguistic tasks where semantic information is essential for providing accurate and efficient performance.

analyses performed in disambiguating systems and sentiment analysis.^{4,5} While humans are able to identify the content and semantics conveyed in written texts in an artlessly manner, automatic classification methods fail in several aspects mainly because the lack of a general representation of world knowledge. Currently, the automatic categorization of texts stands as one of the most studied applications in information sciences. In recent years, multidisciplinary concepts have been applied to assist the classification, including those based on physical approaches.^{3,6–9}

An important sub-area of the research performed in text classification is the detection of subtopics in documents.^{10,11} This task is not only useful to analyze the set of subjects approached in the document but also useful to assist computational applications, such as automatic text summarization¹² and text recommendation.¹³ Several approaches have been proposed to tackle the subtopic identification problem. The most common feature employed to cluster together related sentences/paragraphs is the frequency of shared words.^{14,15} While most of the works on this area consider only the presence/absence of specific words in sentences/paragraphs, only a few studies have considered the organization of words to classify documents.¹⁶ Specifically, in the current paper, we approach the text segmentation problem using the organization of words as a relevant feature for the task. The non-trivial relationships between concepts are modeled via novel networked representations.

Complex networks have been used to analyze a myriad of real systems,¹⁷ including several features of the human language.¹⁸ A well-known model for text representation is the so-called word adjacency (co-occurrence) model, where two words are linked if they appear as neighbors in the text. Though seemingly simple, this model is able to capture

I. INTRODUCTION

With the ever-increasing amount of information available on-line, the classification of texts has established itself as one of the most relevant applications supporting the development of efficient and informative search engines.¹ As the correct categorization of texts is of chief importance for the success of any search engine, this task has been extensively studied, with significant success in some scenarios.^{2,3} However, the comprehension of texts in a human-like fashion remains a challenge, as it is the case with semantical

^{a)}Electronic addresses: diego.rafael@gmail.com and diego@icmc.usp.br

authors' styles,^{19–21} textual complexity,^{8,22} and many other textual aspects.^{18,23,24} A similar model, referred to as syntactical network, takes into account the syntactical representation of texts by connecting words syntactically related. Syntactical networks shared topological properties of other real world networks, including both scale-free and small-world behaviors. These networks have been useful, for example, to capture language-dependent features.²⁵ Unlike traditional models, in this paper, we extend traditional text representations to capture the semantical features so that words are connected if they are related semantically. Here, we take the view that a given subtopic is characterized by a set of words which are internally connected, with a few external relationships with words belonging to other subjects. As we shall show, this view can be straightforwardly translated into the concept of community structure, where each semantical structure can be regarded as a different network community.²⁶ The proposed framework was evaluated in a set of Wikipedia articles, which are tagged according to their subjects. Interestingly, the network approach turned out to be more accurate than traditional methods when applied to several studied datasets, suggesting that the structure of subjects can be used to improve the task. Taken together, our proposed methods could be useful to analyze the written texts in a multilevel networked fashion, as revealed by the topological properties obtained from traditional word adjacency models and networks of networks in higher hierarchies via community analysis.

This manuscript is organized as follows. In Section II, we present the proposed network representation to tackle the subtopic identification task. In Section III, we present the dataset employed for the task. The results obtained are then shown in Section IV. Finally, in Section V, we conclude this paper by presenting perspectives for the further research on related areas.

II. COMPLEX NETWORK APPROACH

The approach we propose here for segmenting texts according to the subjects relies upon a networked representation of texts. In this section, we also detail the method employed for identifying network communities and for the proposed methodology requires a prior segmentation of text networks.

A. Representing texts as networks

A well-known representation of texts as networks is the word-adjacency model, where each different word is a node and links are established between adjacent words.^{27,28} This model, which can be seen as a simplification of the so-called syntactical networks,²⁵ has been used successfully to identify styles in the applications related to authorship attribution, language identification, and authenticity verification.^{29–31} The application of this model in clustering analysis is not widespread because traditional word adjacency networks are not organized in communities, if one considers large pieces of texts. In this paper, we advocate that there is a strong relationship between community structures and subjects. For this reason, we modify the traditional word adjacency model to

obtain an improved representation of word interactions in the written texts. More specifically, here we present a threefold extension which considers different strategies of linking non-adjacent words.

Prior to the creation of the model itself, some preprocessing is applied. First, we remove *stop-words*, i.e., the very frequent words conveying no semantic meaning are removed (e.g., *to*, *have*, and *is*). Note that such words may play a pivotal role in style identification; however, in this study, they are not relevant because such words are subject independent. The preprocessing step is also responsible for removing other non-relevant tokens, such as punctuation marks. Because we are interested in representing concepts as network nodes, it is natural to consider that variant forms of the same word become a unique node of the network. To do so, we lemmatize the words so that nouns and verbs are mapped into their singular and infinitive forms.³² To assist this process, we label each word according to their respective parts of speech (POS) in order to solve possible ambiguities. This part of speech labeling is done with the high-accurate maximum entropy model defined in Refs. 33 and 34. Finally, we consider only nouns and verbs, which are the main classes of words conveying semantical meanings.^{35,36} In all proposed models, each remaining word (lemmatized nouns and verbs) becomes a node. Three variations concerning the establishment of links between such words are proposed:

- *Extended co-occurrence (EC) model*: In this model, the edges are established between two words if they are separated by $d = (\omega - 1)$ or less intermediary words in the preprocessed text. For example, if $\omega = 2$ and the preprocessed text comprises the sentence " $w_1 w_2 w_3 w_4$," then the following set of edges is created: $\mathcal{E} = \{w_1 \leftrightarrow w_2, w_2 \leftrightarrow w_3, w_3 \leftrightarrow w_4, w_1 \leftrightarrow w_3, w_2 \leftrightarrow w_4\}$. If $w = 1$, the traditional co-occurrence (word adjacency) model is recovered, because only adjacent words are connected. Note that even words appearing in distinct (adjacent) paragraphs are connected provided that they appear separated by d or less intermediary words.
- *Paragraph-based (PB) model*: In this model, we link in a clique the words belonging to the same paragraph. We disregard, however, the edges linking words separated by more than d intermediary words in the preprocessing text. This method relies on the premise that paragraphs represent the fundamental sub-structure in which a text is organized, whose words are semantically related.³⁷
- *Adapted paragraph-based (APB) model*: The PB model does not take into account the fact that words may co-occur in the same paragraph just by chance, as it is the case of very frequent words. To consider only significant links in the PB model, we test the statistical significance of co-occurrences with regard to random, shuffled texts.³⁸ Given two words v_i and v_j , an edge is created only if the frequency of co-occurrences (k) (i.e., the number of paragraphs in which v_i and v_j co-occur) is much higher than the same value expected in a null model. The significance of the frequency k can be computed using the quantity $p(k)$, which quantifies the probability of two words to appear in the same paragraph considering the null model.

To compute the probability $p(k)$, let n_1 and n_2 be the number of distinct partitions in which v_i and v_j occur, respectively. The distribution of probability for k , the number of co-occurrences between v_i and v_j , can be computed as

$$p(k) = \frac{(N; k, n_1 - k, n_2 - k)}{(N; n_1)(N; n_2)},$$

where

$$(x; y_1, \dots, y_n) \equiv \frac{x!}{x_1! \dots x_n!} \frac{1}{(x - y_1 - \dots - y_n)!}.$$

The above expression for $p(k)$ can be rewritten in a more convenient way, using the notation,

$$\{a\}_b \equiv \prod_{i=0}^{b-1} (a - i), \quad (1)$$

which is adopted for $a \geq b$. In this case, the likelihood $p(k)$ can be written as

$$\begin{aligned} p(k) &= \frac{\{n_1\}_k \{n_2\}_k \{N - n_1\}_{n_2 - k}}{\{N\}_{n_2} \{k\}_k} \\ &= \frac{\{n_1\}_k \{n_2\}_k \{N - n_1\}_{n_2 - k}}{\{N\}_{n_2 - k} \{N - n_2 + k\}_k \{k\}_k} \\ &= \prod_{j=0}^{n_2 - k - 1} \left[\frac{N - j - n_1}{N - j} \right] \prod_{j=0}^{k-1} \frac{(n_1 - j)(n_2 - j)}{(N - n_2 + k - j)(k - j)}. \end{aligned} \quad (2)$$

If in a given text the number of co-occurrences of two words is r , the p -value p associated to r can be computed as

$$p(k \geq r) = \sum_{k \geq r} p(k), \quad (3)$$

where $p(r)$ is computed according to Equation (2). Now, using Equation (3), the most significant edges can be selected.

Figure 1 illustrates the topology obtained for the three proposed models representing a text with four paragraphs from Wikipedia.³⁹ Note that the structure of communities depends on the method chosen to create the network. Especially, a lower modularity has been found for the APB model in this example and for this reason the organization in communities is not so clear. As we shall show the identification of communities of extreme importance for identifying accurately the subjects.

B. From network communities to text subjects

The first step for clustering texts according to the subjects concerns the computation of network communities, i.e., a region of nodes with several intra-edges (i.e., edges inside the community) and a few inter-edges (i.e., edges leaving the community). Methods for finding network communities have been applied in several applications,^{41–43} including text

analysis.⁴⁴ The quality of partitions obtained from community structure detection methods can be obtained from the modularity Q , which is defined as

$$Q = \frac{1}{2M} \sum_i \sum_j \left(a_{ij} - \frac{k_i k_j}{2M} \right) \delta(c_i, c_j), \quad (4)$$

where a_{ij} denotes an element of the adjacent matrix (i.e., $a_{ij} = 1$ if i and j are linked and $a_{ij} = 0$, otherwise), c_i represents the membership of the i -th node, $k_i = \sum_j a_{ij}$ is the node degree, $M = 1/2 \sum_i \sum_j a_{ij}$ is the total number of edges in the network, and $\delta(x, y)$ is the Kronecker's delta. Several algorithms use the modularity to assist the identification of partitions in networks. Note that the modularity defined in Equation (4) quantifies the difference between the actual number of intra-links (i.e., the links in the same community, $a_{ij} \delta(c_i, c_j)$) and the expected number of intra-links (i.e., the number of links in the same community of a random network, $(k_i k_j / 2M) \delta(c_i, c_j)$). Here, we use a simple yet efficient approach devised in Ref. 40, where the authors define a greedy optimization strategy for Q . More specifically, to devise a greedy optimization, the authors define the quantities

$$e_{ij} = \frac{1}{2M} \sum_v \sum_w a_{vw} \delta(c_v, i) \delta(c_w, j), \quad (5)$$

which represents the fraction of edges that link nodes in community i and j , and

$$\alpha_i = \frac{1}{2M} \sum_v k_v \delta(c_v, i), \quad (6)$$

which denotes the fraction of ends of edges that are linked to nodes in community i . The authors redefine Q in Equation (4) by replacing $\delta(c_v, c_w)$ to $\sum_i \delta(c_v, i) \delta(c_w, i)$

$$\begin{aligned} Q &= \frac{1}{2M} \sum_v \sum_w \left[a_{vw} - \frac{k_v k_w}{2M} \right] \sum_i \delta(c_v, i) \delta(c_w, i) \\ &= \sum_i \left[\frac{1}{2M} \sum_v \sum_w a_{vw} \delta(c_v, i) \delta(c_w, i) \right. \\ &\quad \left. - \frac{1}{2M} \sum_v k_v \delta(c_v, i) \frac{1}{2M} \sum_w k_w \delta(c_w, i) \right] \\ &= \sum_i (e_{ii} - \alpha_i^2). \end{aligned} \quad (7)$$

Using the modularity obtained in Equation (7), it is possible to detect communities using an agglomerative approach. First, each node is associated to a distinct community. In other words, the partition initially comprises only singleton clusters. Then, nodes are joined to optimize the variation of modularity, ΔQ_{ij} . Thus, after the first agglomeration, a multi-graph is defined so that communities are represented as a single node in a new adjacency matrix with elements $a'_{ij} = 2M e_{ij}$. More specifically, the operation of joining nodes is optimized by noting that $\Delta Q_{ij} > 0$ only if communities i and j are adjacent. The implementation details are provided in Ref. 40. Note that our approach does not rely on any

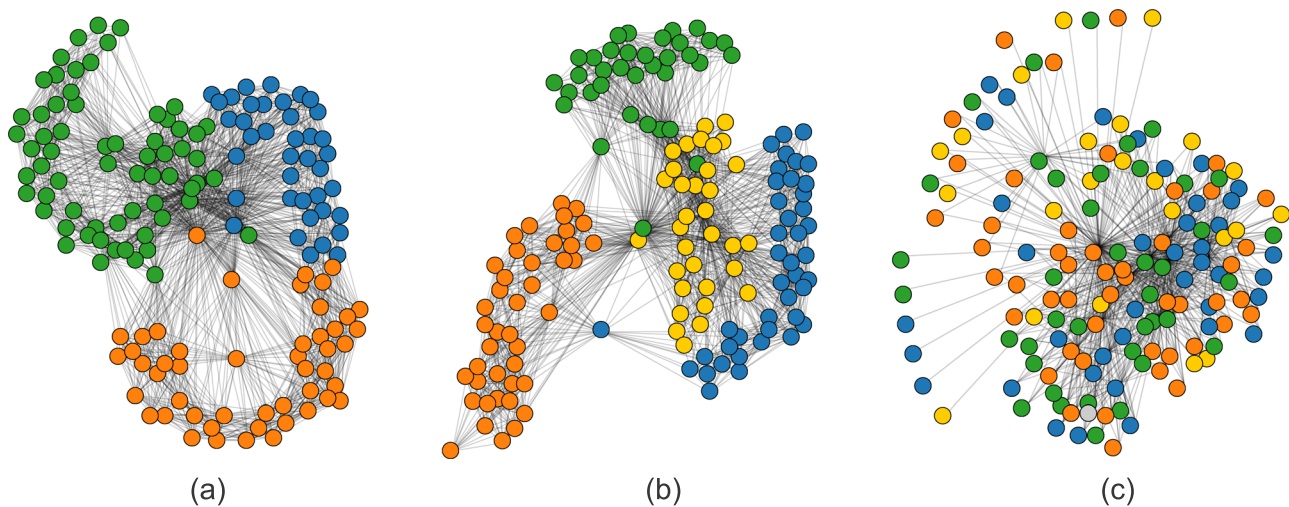


FIG. 1. Example of networks obtained from the following models: (a) EC, (b) PB, and (c) APB. The colors represent in this case the community to which each node belongs. To obtain the partitions, we used the fast greedy algorithm.⁴⁰ Note that the organization of nodes in communities varies from model to model. While in (a) and (b) the organization in communities is clear, in (c) the modular organization is much weaker.

specific community identification method. We have specifically chosen the fast-greedy method because, in preliminary experiments, it outperformed other similar algorithms, such as the multi-level approach devised in Ref. 45 (result not shown).

Given a partition of the network established by the community detection method, we devised the following approach for clustering the text in n_s distinct subjects. Let $\mathcal{C} = \{c_1, c_2, \dots\}$ and $\Pi = \{\pi_1, \pi_2, \dots\}$ be the set of communities in the network and the set of paragraphs in the text, respectively. If the obtained number of communities (n_c) is equal to the expected number of subjects n_s , then we assign the label c_i to the word that corresponds to node i in the text. As a consequence, each paragraph is represented by a set of labels $\mathcal{L}(\pi_j) = \{l_1^{(\pi_j)}, l_2^{(\pi_j)}, \dots\}$, where $l_i^{(\pi_j)} \in \mathcal{C}$ is the label associated to the i -th word of the j -th paragraph (π_j). The number of occurrences of each label in each paragraph is defined as

$$f(c_i, \pi_j) = \sum_{l \in \mathcal{L}(\pi_j)} \delta(c_i, l). \quad (8)$$

Thus, the subject associated to the paragraph π_j is

$$\tilde{s}(\pi_j) = \arg \max_{c_i \in \mathcal{C}} f(c_i, \pi_j), \quad (9)$$

i.e., the most frequent label is chosen to represent the subject of each paragraph. To illustrate the application of our method, Figure 2 shows the communities obtained in a text about cars, whose first paragraph approaches the definition and invention of cars, and the remaining paragraphs present their parts.

The expression in Equation (9) is suitable to be applied only when $n_c = n_s$. Whenever the number of expected subjects is lower than the number of network communities, we adopt a strategy to reduce the quantity of communities found. To do so, the following algorithm is applied:

Data: n_c , the number of communities, and n_s , the number of subjects.

Result: A match between communities and subjects is established.

```

while ( $n_c > n_s$ ) do
     $c_k$  = label of the community with the largest overlapping
    region;
    erase all nodes from  $c_k$ ;
    detect again the community structure of the network
    formed of the remaining nodes;
    update  $n_c$ ;
end

```

According to the above algorithm, if the most overlapping community is to be estimated, a measure to quantify the degree of overlapping must be defined. In the present work, we defined an overlapping index, σ , which is computed for each community. To compute $\sigma(c_i)$, we first recover all paragraphs whose most frequent label is the one associated to community c_i . These paragraphs are represented by the subset $\Pi_D(c_i) = \{\pi \in \Pi \mid \tilde{s}(\pi) = c_i\}$. Next, we count how many words in $\Pi_D(c_i)$ are associated with the community c_i . The overlapping is then defined as the amount of words in $\Pi_D(c_i)$ which are associated to a community $c_j \neq c_i$. Mathematically, the overlapping is defined as

$$\sigma(c_i) = 1 - \sum_{\pi \in \Pi_D(c_i)} \sum_{l \in \mathcal{L}(\pi)} \delta(c_i, l). \quad (10)$$

To evaluate the performance of the methods, we employed the following methodology. Let $s(\pi_i)$ be the subject associated to the i -th paragraph according to Wikipedia. Here, we represent each different subtopic as a integer number, so that $s(\pi_i) \in [1, n_s]$. Let $\tilde{s}(\pi_i) \in [1, n_s]$ be the label associated to the i -th paragraph according to a specific clustering method, as defined in Equation (9). To quantify the similarity of two sets $S = \{s(\pi_1), s(\pi_2), \dots\}$ and $\tilde{S} = \{\tilde{s}(\pi_1), \tilde{s}(\pi_2), \dots\}$, it is necessary to consider all combinations of labels permutations either on S or \tilde{S} , because the same partition may be defined with distinct labels. For example, if $n_s = 2$, $S = \{1, 1, 2, 2\}$,

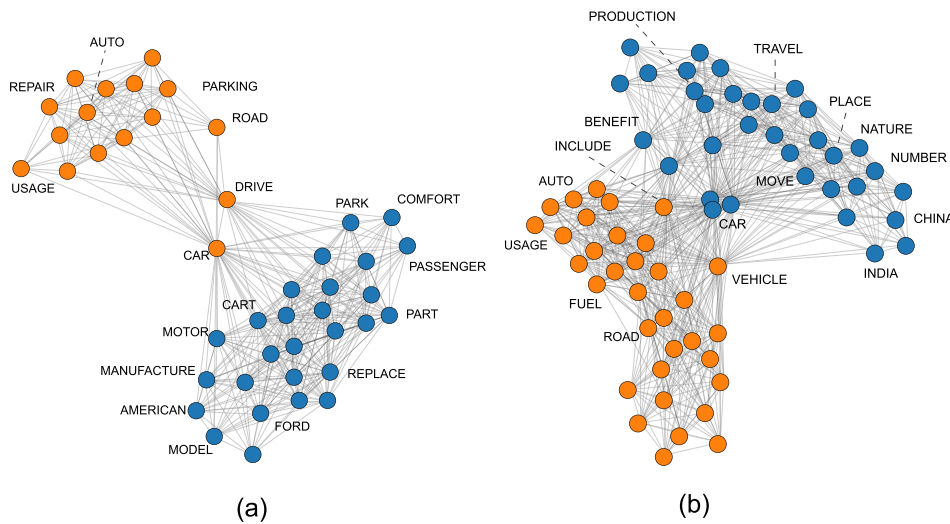


FIG. 2. Example of community obtained with the following models: (a) EC and (b) PB. As expected, most of the words belong to a unique subject, while a few words lie at the overlapping region. The visualization of communities obtained with the APB model is not as good as those obtained for both EC and PB methods.

and $\tilde{S} = \{2, 2, 1, 1\}$, both partitions are the same, even if a straightforward comparison (element by element) yields a zero similarity. To account for distinct labeling in the evaluation, we define the operator \mathcal{P} , which maps a sequence of labels to all possible combinations. Thus, if $S = \{1, 1, 2, 2\}$, then

$$\mathcal{P}(S) = \{\{1, 1, 2, 2\}, \{2, 2, 1, 1\}\}.$$

Equivalently, the application of the \mathcal{P} to S yields two components

$$\mathcal{P}(S, 1) = \{1, 1, 2, 2\} \text{ and } \mathcal{P}(S, 2) = (2, 2, 1, 1).$$

The accuracy rate, Γ , is then defined as

$$\Gamma = \max_i \mathcal{H}(S, \mathcal{P}(\tilde{S}, i)), \quad (11)$$

where $\mathcal{H}(X, Y)$ is the operator that compares the similarity between two subsets $X = \{x_1, x_2, \dots\}$ and $Y = \{y_1, y_2, \dots\}$ and is mathematically defined as

$$\mathcal{H}(X, Y) = \sum_i \delta(x_i, y_i). \quad (12)$$

III. DATABASE

The artificial dataset we created to evaluate the methods is formed from the paragraphs extracted from Wikipedia articles. The selected articles can be classified in 8 distinct topics and 5 distinct subtopics:

- (1) **Actors:** Jack Nicholson, Johnny Depp, Robert De Niro, Robert Downey, Jr., and Tom Hanks.
- (2) **Cities:** Barcelona (Spain), Budapest (Hungary), London (United Kingdom), Prague (Czech Republic), and Rome (Italy).
- (3) **Soccer players:** Diego Maradona (Argentina), Lionel Messi (Argentina), Neymar, Jr. (Brazil), Pelé (Brazil), and Robben (Netherlands).
- (4) **Animals:** Bird, cat, dog, lion, and snake.
- (5) **Food:** Bean, cake, ice cream, pasta, and rice.
- (6) **Music:** Classical, funk, jazz, rock, and pop.

- (7) **Scientists:** Albert Einstein, Gottfried Leibniz, Linus Pauling, Santos Dumont, and Alan Turing.
- (8) **Sports:** Football, basketball, golf, soccer, and swimming.

To construct the artificial texts, we randomly selected paragraphs from the articles using two parameters: n_s , the total number of subtopics addressed in the artificial text; and n_p , the number of paragraphs per subtopic. For each pair (n_s, n_p) , we compiled a total of 200 documents. To create a dataset with distinct granularity of subtopics, the random selection of subtopics was performed in a two-fold way. In the dataset comprising coarse-grained subtopics, hereafter referred to as CGS dataset, each artificial text comprises n_s subtopics of *distinct* topics. Differently, in the dataset encompassing fine-grained subtopics, hereafter referred to as FGS dataset, the texts are made up of distinct subtopics of the *same* major topic.

A drawback associated to the creation of the described artificial dataset is that textual self-consistency may be compromised when paragraphs are combined. Conversely, the adoption of this procedure allows us to control several text parameters, which in turn makes possible the analysis of the influence of text properties on the performance of our methods. As we shall show, the adoption of this strategy allowed the identification of several interesting patterns, including the dependence of performance on subject granularity and text length.

IV. RESULTS

In this section, we analyze the statistical properties of the proposed models in terms of their organization in communities. We also evaluate the performance of our model in the artificial dataset and compare with more simple models that do not rely on any networked information.

A. Modularity analysis

The models we propose to cluster topics in texts rely on the ability of a network to organize itself in a modular way. For this reason, we study how the modularity of networks depends on the models parameters. The unique parameter that may affect networks modularity in our models is ω ,

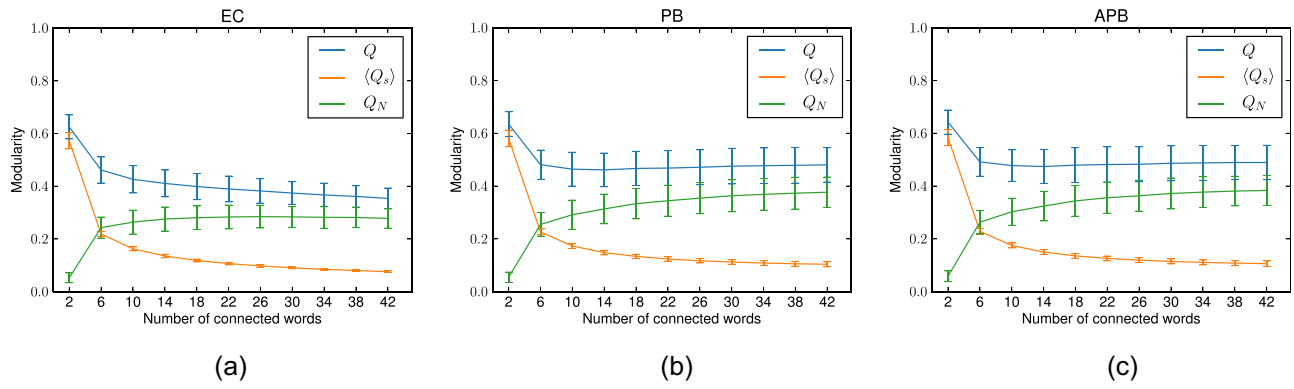


FIG. 3. Example of modularity evolution using the proposed text representations (EC, PB, and APB). The results were obtained in both CGS and FGS datasets (see Section III). Note that the normalized modularity Q_N does not display a significant increase for $\omega > 20$.

which amounts to the distance of links in texts (see definition in the description of the EC model in Section II). Note that the distance ω controls the total number of edges of the network, so a comparison of networks modularity for distinct values of ω is equivalent to comparing networks with distinct average degrees. Because the average degree plays an important role on the computation of the modularity,²⁶ we defined here a normalized modularity Q_N that only takes into account the organization of the network in modules and is not biased towards dense networks. The normalized

modularity Q_N of a given network with modularity Q is computed as

$$Q_N = Q - \langle Q_s \rangle, \quad (13)$$

where Q_s is the average modularity obtained in 30 equivalent random networks with the same number of nodes and edges of the original network.

In Figure 3, we show the values of Q , Q_N , and Q_s in the dataset FGS. In the EC model, high values of modularity Q

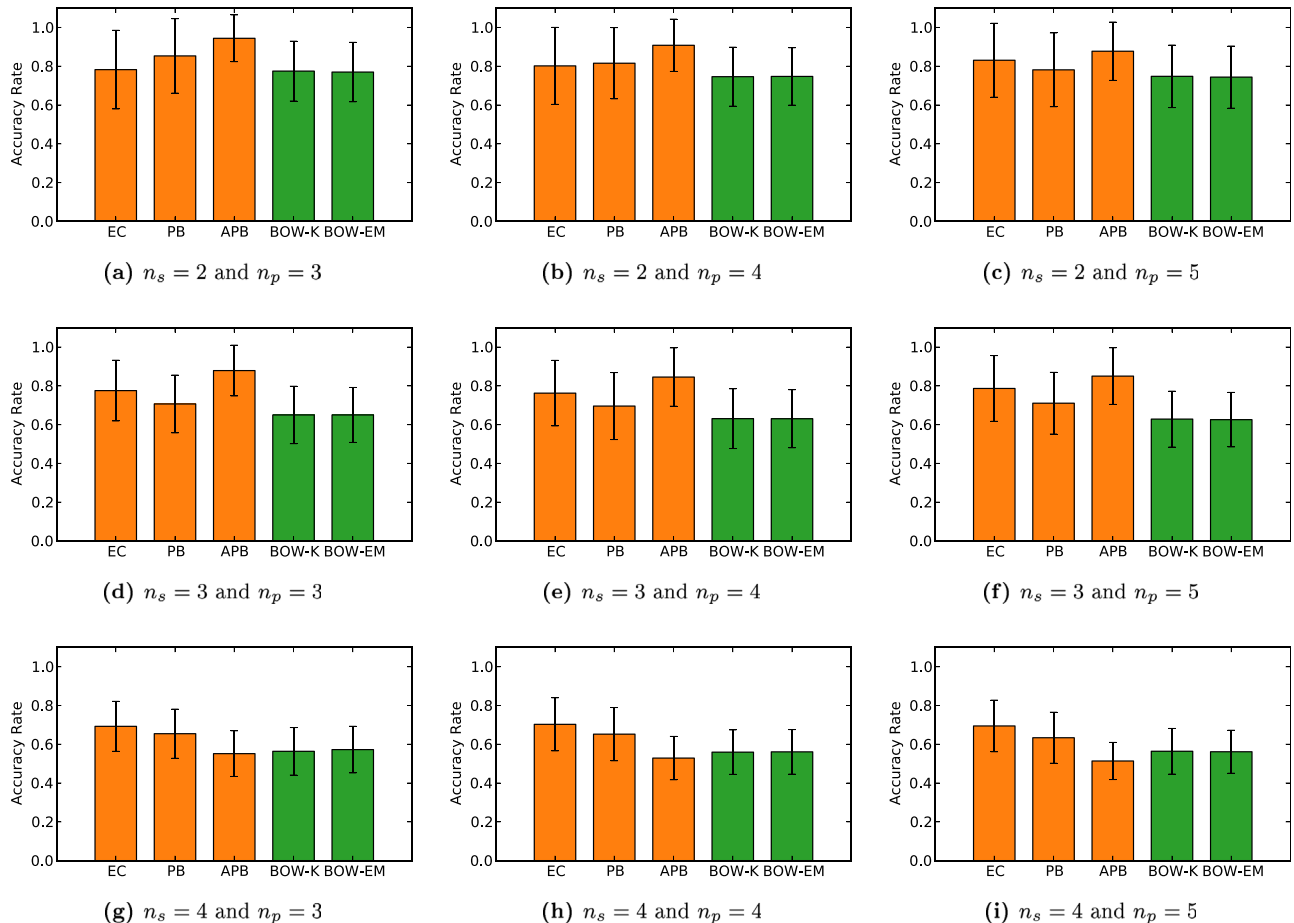


FIG. 4. Performance in segmenting subjects in texts with subtopics on the same major topic. The parameters employed to generate the dataset (n_s , the number of subtopics, and n_p , the number of paragraphs per subtopic) are shown in the figure. Note that, in most of the cases, the best performance is obtained with the APB approach.

are found for low-values of ω . To the best of our knowledge, this is the first time that a modular structure is found in a traditional word adjacency network (i.e., the EC model with $\omega = 1$) in a relatively small network. As ω takes higher values and the network becomes denser, the modularity Q decreases. In a similar way, the average modularity $\langle Q_S \rangle$ obtained in equivalent random networks also decreases when ω increases. A different behavior arises for the normalized modularity Q_N , as shown in Figure 3(a). The modularity $\langle Q_S \rangle$ initially takes a low value, reaching its maximum when $\omega = \omega_{max} = 20$. Note that when $\omega > \omega_{max}$, there is no significant gain in modularity Q_N . A similar behavior appears in the other two considered models (PB and APB) as the normalized modularity becomes almost constant for $\omega > 20$. Because in all models the normalized modularity takes high values for $\omega = 20$, we have chosen this value to construct the networks for the purpose of clustering topics. Even though a higher value of normalized modularity was found for $\omega > 20$ in Figures 3(b) and 3(c), we decided not to use larger values of ω in these models because only a minor improvement in the quality is obtained for $\omega > 20$.

In auxiliary experiments, we have verified whether the use of other methods for community detection is able to provide improved values of Q . The methods evaluated were the fastgreedy, leading eigenvector, walktrap, infomap, edge betweenness, multilevel, label propagation, and spinglass.²⁶

The obtained results are shown in Figures S1 and S2 of the [supplementary material](#). Even though some methods performed better than others in particular cases, we have found no significant difference in performance when analyzing the best methods. For this reason, we restricted the discussion below to a single method.

B. Performance analysis

To evaluate the performance of the proposed network-based methods, we evaluate the accuracy of the generated partitions in the CGS and FGS datasets presented in Section III. We have created two additional methods based on linguistic attributes to compare the performance of networked and non-networked methods. Both methods are based on a *bag-of-words* strategy,¹ where each paragraph is represented by the frequency of appearance of its words. To cluster the paragraphs in distinct subjects, we used two traditional clustering methods: k-means⁴⁶ and Expectation Maximization.⁴⁷ In our results, the bag-of-words strategy based on the k-means and Expectation Maximization algorithms are referred to as BOW-K and BOW-EM, respectively.

In Figure 4, we show the results obtained in the dataset comprising texts whose subtopics belong to distinct major topics (CGS dataset). We classified the dataset in terms of the number of distinct subtopics in each text (n_s) and the total

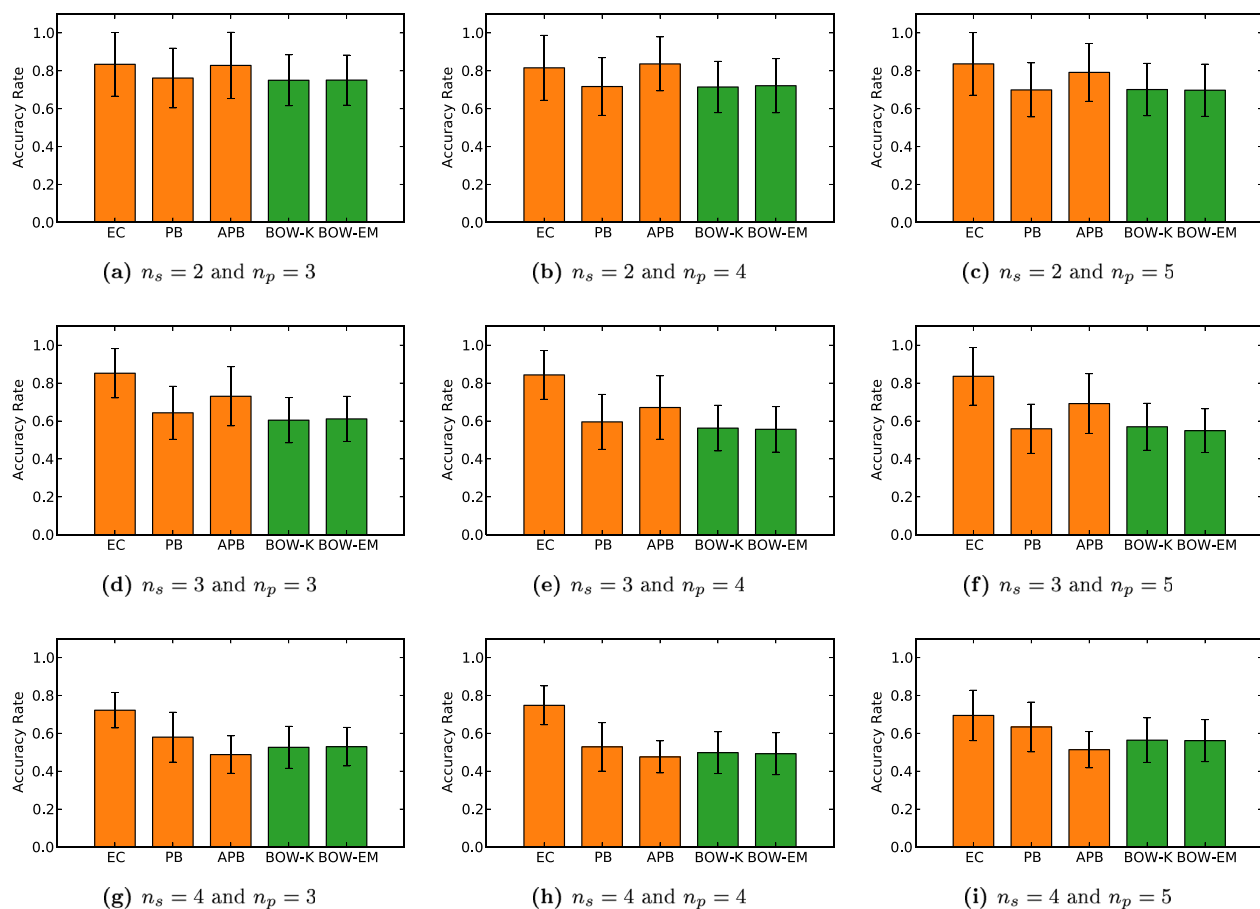


FIG. 5. Performance obtained in the FGS dataset. The parameters employed to generate the dataset (n_s , the number of subtopics, and n_p , the number of paragraphs per subtopic) are shown in the figure. In most of the cases, the best performance was obtained with the EC approach.

number of paragraphs per subtopic (n_p) (see Section III). We first note that, in all cases, at least one networked approach outperformed both BOW-K and BOW-EM approaches. In some cases, the improvement in performance is much clear (see Figures 4(d)–4(f)), while in others the difference in performance is lower. When comparing all three networked approaches, the APB strategy, in average, performs better than others when the number of subtopics is $n_s \leq 3$. For $n_s = 4$, both EC and PB strategies outperform the APB method.

In Figure 5, we show the performance obtained in the dataset comprising texts with subtopics belonging to the same major topic (FGS dataset). When one compares the networked approaches with BOW-K and BOW-EM, we note again that at least one networked approach outperformed both non-networked strategies. This result confirms that the networked method seems to be useful especially when the subtopics are not very different from each other, which corresponds to the scenario found in most real documents. The relevance of our proposed methods is specially clear when $n_s = 3$. A systematic comparison of all three networked methods reveals that the average performance depends on specific parameters of the dataset. In most of the cases, however, the best average performance was obtained with the EC method. The APB method also displayed high levels of accuracy, with no significant difference of performance in comparison with EC in most of the studied cases.

All in all, we have shown that networked approaches tend to perform better than traditional techniques that do not rely on a networked representation of texts. Interestingly, a larger gain in performance was obtained in the FGS dataset, where the difference in subtopics is much more subtle. This result suggests that the proposed networked approaches are potentially more suitable to analyze real texts, as one expects that the changes in subjects in a given text are much more subtle than variations of subjects across distinct texts. Another interesting finding concerns the variations of performance in distinct datasets. Our results revealed that there is no unique networked approach able to outperform strategies in all studied cases. However, we have observed that, in general, both EC and APB methods perform better than the AP approach and, for this reason, they could be tried in real applications.

V. CONCLUSION

In this paper, we proposed a method to find subtopics in written texts using the structure of communities obtained in word networks. Even though texts are written in a modular and hierarchical manner,⁴⁸ such a feature is hidden in traditional networked text models, as it is the case of syntactical and word adjacency networks. In order to capture the topic structure of a text, we devised three methods to link words appearing in the same semantical context, whose length is established via parameterization. In special, the proposed method presents a very low dependence with the specificities of the analyzed language, as the main linguistic resource used was the part-of-speech (POS) tagger. In preliminary experiments, we have found that the modular organization is

optimized when one considers a context comprising 20 words. Above this threshold, the gain in modularity was found to be not significant. We applied our methods in a subset of articles from Wikipedia with two granularity levels. Interestingly, in all considered datasets, at least one of the proposed networked methods outperformed traditional bag-of-word methods. As a proof of principle, we have shown that the information of network connectivity hidden in texts might be used to unveil their semantical organization, which in turn might be useful to improve the applications relying on the accurate characterization of textual semantics.

In future works, we intend to use the proposed characterization to study related problems in text analysis. The suggested high-level representation could be employed in a straightforward manner to recognize styles in texts, as one expects that distinct writing styles may approach different subjects in a very particular way. As a consequence, the proposed networked representations could be used for identifying authors in disputed documents. We also intend to extend our methods to make it suitable to analyze very large documents. In this context, an important issue concerns the accurate choice of values for the context length, which plays an important role in the performance. Our methods could also be combined with other traditional networked representation to improve the characterization of related systems. Thus, a general framework could be created to study the properties of written texts in a multi-level way, in order to capture the topological properties of networks formed of simple (e.g., words) and more complex structures (e.g., communities).

SUPPLEMENTARY MATERIAL

See [supplementary material](#) for additional modularity results obtained with traditional community detection methods.

ACKNOWLEDGMENTS

Henrique F. de Arruda thanks Federal Agency for Support and Evaluation of Graduate Education (CAPES-Brazil) for the financial support. Diego R. Amancio acknowledges São Paulo Research Foundation (FAPESP) (Grant No. 2014/20830-0) for the financial support. Luciano da Fontoura Costa is grateful to CNPq (Grant No. 307333/2013-2), FAPESP (Grant No. 11/50761-2), and NAP-PRP-USP for sponsorship.

¹C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing* (MIT Press, Cambridge, MA, USA, 1999).

²C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in *Mining Text Data* (Springer, US, 2012), pp. 163–222.

³J. P. Herrera and P. A. Pury, "Statistical keyword detection in literary corpora," *Eur. Phys. J. B* **63**, 135–146 (2008).

⁴S. A. Golder and M. W. Macy, "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures," *Science* **333**, 1878–1881 (2011).

⁵R. Navigli, "Word sense disambiguation: A survey," *ACM Comput. Surv.* **41**, 1–69 (2009).

⁶D. G. Hernández, D. H. Zanette, and I. Samengo, "Information-theoretical analysis of the statistical dependencies among three variables: Applications to written language," *Phys. Rev. E* **92**, 022813 (2015).

- ⁷D. R. Amancio, O. N. Oliveira, Jr., and L. d. F. Costa, "Three-feature model to reproduce the topology of citation networks and the effects from authors visibility on their h-index," *J. Informetrics* **6**, 427–434 (2012).
- ⁸H. Liu and C. Xu, "Can syntactic networks indicate morphological complexity of a language?," *EPL* **93**, 28005 (2011).
- ⁹A. Delanoe and S. Galam, "Modeling a controversy in the press: The case of abnormal bee deaths," *Physica A* **402**, 93–103 (2014).
- ¹⁰J. C. Reynar, "Statistical models for topic segmentation," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, ACL'99* (Association for Computational Linguistics, Stroudsburg, PA, USA, 1999), pp. 357–364.
- ¹¹A. Lancichinetti, M. I. Siler, J. Wang, D. Acuna, K. Kording, and L. A. N. Amaral, "High-reproducibility and high-accuracy method for automated topic classification," *Phys. Rev. X* **5**, 011007 (2015).
- ¹²S. Gong, Y. Qu, and S. Tian, "Subtopic-based multi-documents summarization," in *3rd International Joint Conference on Computational Science and Optimization* (2010), Vol. 2, pp. 382–386.
- ¹³G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin, "Incorporating contextual information in recommender systems using a multidimensional approach," *ACM Trans. Inf. Syst.* **23**, 103–145 (2005).
- ¹⁴C. Clifton, R. Cooley, and J. Rennie, "Topcat: Data mining for topic identification in a text corpus," *IEEE Trans. Knowl. Data Eng.* **16**, 949–964 (2004).
- ¹⁵D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
- ¹⁶D. R. Amancio, "A complex network approach to stylometry," *PLoS One* **10**, e0136076 (2015).
- ¹⁷L. d. F. Costa, O. N. Oliveira, Jr., G. Travieso, F. A. Rodrigues, P. R. V. Boas, L. Antiquiera, M. P. Viana, and L. E. C. Rocha, "Analyzing and modeling real-world phenomena with complex networks: A survey of applications," *Adv. Phys.* **60**, 329–412 (2011).
- ¹⁸J. Cong and H. Liu, "Approaching human language with complex networks," *Phys. Life Rev.* **11**, 598–618 (2014).
- ¹⁹D. R. Amancio, E. G. Altmann, O. N. Oliveira, Jr., and L. d. F. Costa, "Comparing intermittency and network measurements of words and their dependence on authorship," *New J. Phys.* **13**, 123024 (2011).
- ²⁰D. R. Amancio, "Authorship recognition via fluctuation analysis of network topology and word intermittency," *J. Stat. Mech.: Theory Exp.* **2015**, P03005.
- ²¹A. Mehri, A. H. Darooneh, and A. Shariati, "The complex networks approach for authorship attribution of books," *Physica A* **391**, 2429–2437 (2012).
- ²²D. R. Amancio, S. M. Aluisio, O. N. Oliveira, Jr., and L. d. F. Costa, "Complex networks analysis of language complexity," *EPL* **100**, 58002 (2012).
- ²³A. P. Masucci, A. Kalampokis, V. M. Eguíluz, and E. Hernández-García, "Wikipedia information flow analysis reveals the scale-free architecture of the semantic space," *PLoS One* **6**, e17333 (2011).
- ²⁴A. P. Masucci and G. J. Rodgers, "Differences between normal and shuffled texts: Structural properties of weighted networks," *Adv. Complex Syst.* **12**, 113–129 (2009).
- ²⁵R. F. Cancho, R. V. Solé, and R. Köhler, "Patterns in syntactic dependency networks," *Phys. Rev. E* **69**, 051915 (2004).
- ²⁶S. Fortunato, "Community detection in graphs," *Phys. Rep.* **486**, 75–174 (2010).
- ²⁷R. F. Cancho and R. V. Solé, "The small world of human language," *Proc. R. Soc. London, Ser. B* **268**, 2261–2265 (2001).
- ²⁸D. R. Amancio, L. Antiquiera, T. A. S. Pardo, L. d. F. Costa, O. N. Oliveira, Jr., and M. G. V. Nunes, "Complex networks analysis of manual and machine translations," *Int. J. Mod. Phys. C* **19**, 583–598 (2008).
- ²⁹D. R. Amancio, "Comparing the topological properties of real and artificially generated scientific manuscripts," *Scientometrics* **105**, 1763–1779 (2015).
- ³⁰M. A. Montemurro and D. H. Zanette, "Keywords and co-occurrence patterns in the voynich manuscript: An information-theoretic analysis," *PLoS One* **8**, e66344 (2013).
- ³¹D. R. Amancio, E. G. Altmann, D. Rybski, O. N. Oliveira, Jr., and L. d. F. Costa, "Probing the statistical properties of unknown texts: Application to the voynich manuscript," *PLoS One* **8**, e67310 (2013).
- ³²G. A. Miller, "Wordnet: A lexical database for English," *Commun. ACM* **38**, 39–41 (1995).
- ³³A. Ratnaparkhi *et al.*, "A maximum entropy model for part-of-speech tagging," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Philadelphia, USA* (1996), Vol. 1, pp. 133–142.
- ³⁴A. L. Berger, V. J. Della Pietra, and S. A. Della Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguist* **22**(1), 39–71 (1996).
- ³⁵D. Hindle, "Noun classification from predicate-argument structures," in *Proceedings of the 28th Annual Meeting on Association for Computational Linguistics* (Association for Computational Linguistics, 1990), pp. 268–275.
- ³⁶N. Béchet, J. Chauché, V. Prince, and M. Roche, "How to combine text-mining methods to validate induced verb-object relations?," *Comput. Sci. Inf. Syst.* **11**, 133–155 (2014).
- ³⁷J. Véronis, "Hyperlex: Lexical cartography for information retrieval," *Comput. Speech Lang.* **18**, 223–252 (2004).
- ³⁸J. Martínez-Romo, L. Araujo, J. Borge-Holthoefer, A. Arenas, J. A. Capitán, and J. A. Cuesta, "Disentangling categorical relationships through a graph of co-occurrences," *Phys. Rev. E* **84**, 046108 (2011).
- ³⁹See en.wikipedia.org/wiki/Car for obtaining the content used to illustrate the proposed models.
- ⁴⁰A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E* **70**, 066111 (2004).
- ⁴¹A. Arenas, L. Danon, A. Diaz-Guilera, P. M. Gleiser, and R. Guimera, "Community analysis in social networks," *Eur. Phys. J. B* **38**, 373–380 (2004).
- ⁴²R. Guimera, S. Mossa, A. Turttschi, and L. N. Amaral, "The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles," *Proc. Natl. Acad. Sci. U. S. A.* **102**, 7794–7799 (2005).
- ⁴³G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature* **435**, 814–818 (2005).
- ⁴⁴N. Londhe, V. Gopalakrishnan, A. Zhang, H. Q. Ngo, and R. Srihari, "Matching titles with cross title web-search enrichment and community detection," *Proc. VLDB Endowment* **7**, 1167–1178 (2014).
- ⁴⁵V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.: Theory Exp.* **2008**, P10008.
- ⁴⁶T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 881–892 (2002).
- ⁴⁷B. Chai, C. Jia, and J. Yu, "An online expectation maximization algorithm for exploring general structure in massive networks," *Physica A* **438**, 454–468 (2015).
- ⁴⁸E. Alvarez-Lacalle, B. Dorow, J.-P. Eckmann, and E. Moses, "Hierarchical structures induce long-range dynamical correlations in written texts," *Proc. Natl. Acad. Sci. U. S. A.* **103**, 7956–7961 (2006).