Journal of Informetrics 10 (2016) 487–502

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Using network science and text analytics to produce surveys in a scientific topic

Filipi N. Silva [a], Diego R. Amancio [b,*], Maria Bardosova [c], Luciano da F. Costa [a], Osvaldo N. Oliveira Jr. [a]

[a] São Carlos Institute of Physics, University of São Paulo, São Carlos, Brazil
[b] Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, Brazil
[c] Tyndall National Institute, Cork City, Ireland

ARTICLE INFO

Article history:
Received 7 November 2015
Received in revised form 17 March 2016
Accepted 31 March 2016
Available online 22 April 2016

Keywords:
Entropy
Networks
Scientific map
Photonic crystals
Pattern recognition

ABSTRACT

The use of science to understand its own structure is becoming popular, but understanding the organization of knowledge areas is still limited because some patterns are only discoverable with proper computational treatment of large-scale datasets. In this paper, we introduce a framework to combine network-based methodologies and text analytics to construct the taxonomy of science fields. The methodology is illustrated with application to two topics: *complex networks* (CN) and *photonic crystals* (PC). We built citation networks using data from the Web of Science and used a community detection algorithm for partitioning to obtain *science maps* for the two topics. We also created an importance index for text analytics, which is employed to extract keywords that define the communities and, combined with network topology metrics, to generate dendrograms of relatedness among subtopics. Interesting patterns emerging from the analysis included identification of two well-defined communities in PC area, which is consistent with the known existence of two distinct communities of researchers in the area: telecommunication engineers and physicists. With the methodology, it was also possible to assess the interdisciplinary nature and time evolution of subtopics defined by the keywords. The automatic tools described here are potentially useful not only to provide an overview of scientific areas but also to assist scientists in performing systematic research on a specific topic.

1. Introduction

Recent developments in the use of machine learning methods to extract information (and knowledge!) from Big Data have shown that machines are bound to replace humans in various intellectual tasks in the near future, particularly in cases where a lot of information needs to be processed (Bell, Hey, & Szalay, 2009; Craddock, Harwood, Hallinan, & Wipat, 2008; Donovan, 2008). Clear examples of such tasks are facial recognition (Zhao, Chellappa, Phillips, & Rosenfeld, 2003), establishing best routes for cars and passengers (Laporte, 1992), internet search (Lawrence & Giles, 1998), etc. Some authors have even been bold enough to suggest that scientific and technological development is being held back by the limited capacity of humans, especially the memory, to process and interpret the electronic data available (Stone & Lavine, 2014). A specific task in academic work where this limited capacity is readily apparent is in carrying out a survey of any given topic,

* Corresponding author. Tel.: +55 16 992971183.
E-mail address: diego@icmc.usp.br (D.R. Amancio).

http://dx.doi.org/10.1016/j.joi.2016.03.008
1751-1577/

owing to the vast literature to be consulted. The first requirement for a survey, namely to establish a map of knowledge (also known as *science map*) of the field under analysis, demands data-intensive discovery. Surveys normally performed by humans benefit from well-founded techniques to organize scientific literature and information, but little help exists for understanding the knowledge structure on a larger scale. Even experienced researchers find this hard owing to the aforementioned human limited capacity, and there is the additional drawback of bias – even if unintentional – toward the experts' personal preferences. Not surprisingly, modeling the knowledge structure remains an open problem in science with the intricate relationships among the many concepts involved.

In this paper we propose a new framework to assist humans in preparing literature surveys, which consists of the integration of many well-established concepts arising from complex networks (Barabasi & Albert, 1999) that have been proven effective in modeling the organization of knowledge (Börner & Scharnhorst, 2009; Boyack & Klavans, 2014; Boyack, Klavans, & Borner, 2005; Costa, Oliveira, Travieso, Rodrigues, & Villas Boas, 2011; Silva, Rodrigues, Oliveira, & Costa, 2013). Our approach, however, distinguishes itself from previous ones in the literature since network science and text analytics methods are interwoven to generate science maps and taxonomies. More specifically, we build citation networks (Chen & Hicks, 2004; Leicht, Clarkson, Shedden, & Newman, 2007; Menczer, 2004) that serve as the overall framework of a science map, which needs to be complemented with a taxonomy to classify the contents of the map. We adapted the methodologies to extract keywords to complete the science map for two fields, namely "Complex Networks" and "Photonic Crystals". This choice was basically due to the authors of the paper being experts in these fields, which allows for a deeper discussion of the results obtained.

2. Overview of complex networks and text analytics applied to summarization

Because our study deals with two very distinct areas, namely use of complex network methods to analyze scientific literature and text analytics, a brief overview of previous work will be done here for these areas. This overview is by no means exhaustive, particularly as there has been a vast literature in each of these areas; we rather concentrate on work that is directly related to the purpose of our study, which is to provide semi-automated means for assisting authors in surveys of the literature and document summarization techniques (Silva, Viana, Travençolo, & Costa, 2011).

Recent works have used network-based metrics to characterize or quantify relevance and impact of researchers, publications and journals (Ding, Yan, Frazho, & Caverlee, 2009; McKeown et al., 2016; Nykl, Campr, & Ježek, 2015; Yan, Zhai, & Fan, 2013; Zhou, Zeng, Fan, & Di, 2015). For instance, factor analysis was employed to automatically extract the most important papers in citation networks (Chen, 2012). Citation-based networks have been used in various domains, such as modeling the dynamics of knowledge acquisition and dissemination (Amancio, 2015a; Amancio, Nunes, Oliveira, & Costa, 2012c; Börner & Scharnhorst, 2009), enriching and contextualizing information of biological experiments or data (Mullen, Daley, Backx, & Thompson, 2014), and visualizing relationships among scientific fields by constructing science maps (Börner et al., 2012; Boyack et al., 2005; Leydesdorff & Rafols, 2009).

Of particular importance are science maps used as a versatile tool to qualitatively understand how science fields are organized, by e.g. establishing relationships among distinct areas (Boyack et al., 2005; Leydesdorff & Rafols, 2009; Porter & Rafols, 2009; Rosvall & Bergstrom, 2008; Silva et al., 2011; Silva, Travençolo, Viana, & Costa, 2010). Tools have been developed to visualize and interact with scientific maps (Boyack, Wylie, & Davidson, 2002; Silva & Costa, 2011; Silva et al., 2013; van Eck & Waltman, 2010, 2014; Waaijer, van Bochove, & van Eck, 2011), and understand interdisciplinarity (Lariviere, Haustein, & Börner, 2015; Leydesdorff, de Moya-Anegón, & de Nooy, 2015; Leydesdorff, Rafols, & Chen, 2013; Porter & Rafols, 2009; Silva et al., 2013) among scientific journals. In a similar fashion, science maps can also be constructed by using self-organizing maps in which scientific domains are mapped to a 2D space according to a neural network through a Hebbian learning process (Skupin, Biberstine, & Börner, 2013). While science maps are able to provide interesting insights about the overall structure of science, a contextualized taxonomy of its structure is more appropriate to the task of surveying a scientific field. This is because survey papers are conventionally organized in a hierarchical structure, normally comprising chapters, sections, subsections and other forms of text partitions. Establishing such taxonomy, with components and subcomponents hierarchically organized, is not trivial for automated tools (Sebastiani, 2002; Silva et al., 2013), and various procedures have been adopted to classify contents.

Text summarization is a traditional area of text analytics, which has been used to build summaries and taxonomies of text datasets comprising many types of situations, such as tracing the events of disasters using social media (Kedzie, McKeown, & Diaz, 2015), conferences (Shen, Liu, Weng, & Li, 2013) and sports events (Nichols, Mahmud, & Drews, 2012). The main goal with such techniques is to obtain an importance metric (also called *salience*) for terms or sentences. The summary of the content can be constructed by rewriting the text using only terms or sentences presenting high salience, while the taxonomies can be obtained by clustering texts according to the similarities among their most important terms. This can be accomplished through the use of metrics such as cosine similarity (Salton & Buckley, 1988) or semantic-wise similarities (Boyack et al., 2011), as in relationships in the WordNet or word embedding techniques (Levy & Goldberg, 2014). A simple way to obtain the salience of terms is by comparing their relative frequency of appearance inside a document to their frequency of appearance in a larger set of other documents. This is usually referred to as the TF-IDF (Salton & Buckley, 1988) method, which yields good results for sets of large texts. However, the method becomes unreliable when measuring relevance of terms in sets of small texts, since terms tend to appear only a few times for each document, as in paper abstracts and messages of social networking services. Other, more complex, summarization techniques can be used to deal with such

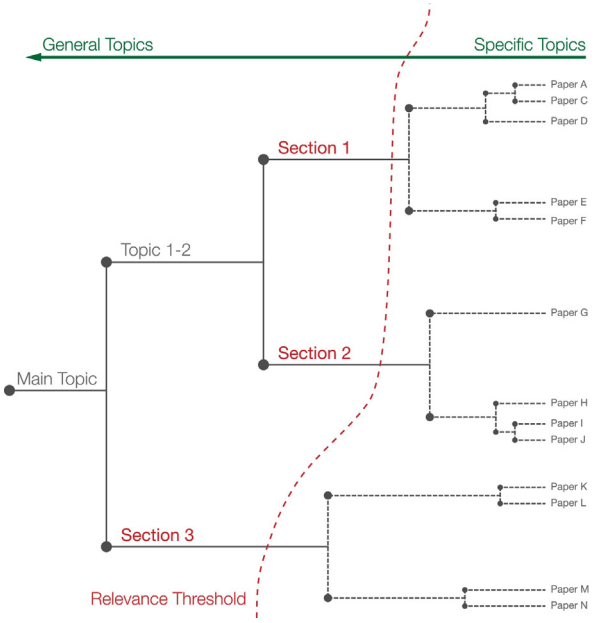

Fig. 1. Example of the structure of an organized scientific survey. Papers are grouped into more general topics which are reflected as sections, subsections, chapters, etc. A threshold of relevance and focus is thus necessary as their content needs to be summarized and cannot retain the full level of detail for each paper.

type of data. Examples are supervised machine learning methods that require a small set of golden summaries used to train a machine to detect important terms. Human readable summaries may be generated from a document or a set of documents (Radev & McKeown, 1998) by using features of low contextual content, such as the average number of words or the number of capitalized words in a sentence (Nenkova & McKeown, 2012).

As an alternative to machine learning methods, topics analysis (Blei, Ng, & Jordan, 2003) has been employed to find important terms (keywords) in a set of documents, such as articles or abstracts (Griffiths & Steyvers, 2004), where terms are projected and clustered according to their presence in a set of documents. This is done by estimating a Markov chain model of topic information along the documents, normally obtained by Gibbs sampling. This technique presents high computational cost, as it requires several iterations to estimate the transitions between words, but it can give good results depending on the size of each document, the number of documents and other properties of the dataset, as studied in depth by Tang, Meng, Nguyen, Mei, and Zhang (2014).

Methods derived from network science have also been used for document summarization. The LexRank technique (Erkan & Radev, 2004) relies upon a network of similarity between sentences to obtain topological centrality measurements, such as eigenvector centrality Newman (2010). The centrality measurements are then used to quantify the salience of terms. In a similar fashion, word adjacency networks were employed to find keywords in a text (Amancio, Nunes, Oliveira, & Costa, 2012b), where salience was obtained from the diversity measurement of nodes (Viana, Travencolo, Tanck, & Costa, 2010) and provided superior results to traditional centrality measurements in networks. Such kind of analysis is advantageous compared to multiple text analytics methods for the same dataset since information provided by network based techniques does not overlap with that provided by traditional text analytics techniques (Amancio, 2015b; Amancio, Oliveira, & Costa, 2012e; Li, Zhou, Luo, & Yang, 2012; Newman & Clauset, 2015; Silva & Amancio, 2012).

3. Methodology

A survey paper is taken here as an organized structure that summarizes information about a scientific field. It must limit the level of detail for each topic by highlighting the most relevant pieces of information while also reducing their redundancy. The hierarchy in a survey comprises concepts that are progressively merged together by their relatedness to build major contextual structures such as subsections and sections, as exemplified in Fig. 1. Topics are hierarchically structured, each of which can represent a set of papers or other scientific works relevant to the area.

To determine the hierarchy of components and subcomponents, we first built citation networks for the fields *Complex Networks* (CN) and *Photonic Crystals* (PC), whose papers were retrieved from the Web of Science (WOS)[1] database using the query terms “complex network” and “photonic crystal” (including the plural variations), respectively. For each retrieved

[1] http://thomsonreuters.com/thomson-reuters-web-of-science/

paper, we extracted the title, abstract, publication year, citation count and list of references. Two citation networks were built (CN and PC) where nodes represent the papers and an edge was established between two papers if one cites the other.

There are many ways to construct citations-based networks. They can be drawn directly from the citation structure, in which two papers are connected if there is a citation between them, resulting in an unweighted directed network. Also used in several studies are co-citation networks (Üsdiken & Pasadeos, 1995; Chen, 2004; Ding et al., 2009; Jenssen, Lægreid, Komorowski, & Hovig, 2001), where documents are connected if they share a citation with at least another document. This procedure leads to a weighted undirected network, and the number of shared documents can be used as a metric of similarity among documents.

For the sake of simplicity, here we opted to use traditional citation networks, but we do not take into account the direction of citation connections. We understand that this information is relevant in several other studies (Chen & Hicks, 2004; Menczer, 2004), but not here because we use citation networks to represent a knowledge relationship structure which is naturally undirected. As an alternative, we also applied the analysis presented in this work to co-citation networks as shown in Supplementary material, and found similar results in the analysis. However, such networks are denser and harder to discuss and visualize.

The citation networks were constructed by first obtaining the vertices from papers returned from the chosen queries for CN and PC in the Web of Science dataset. Next, citation information was used to connect pairs of cited papers where papers that were not present in the initial queries were ignored (even if cited by others). This avoids problems caused by dangling nodes, which can impact the topological analysis employed here, such as community detection.

Citation networks can be transformed into science maps if the most relevant topics and their inter-relationships are identified. In this study, the CN and PC citation networks were embedded in a 3D space using a force-directed method based on the Fruchterman-Reingold algorithm (Fruchterman & Reingold, 1991). The initial configuration had the nodes, treated as particles, uniformly distributed over a 3D space. These nodes were allowed to interact via repulsive forces, with attractive forces being added for the connected nodes. When the energy of the whole system was minimized, the resulting embedding became a graphically appealing projection of the network topology (Bando, Silva, Costa, Silva, & Pimentel-Silva, 2013; Silva et al., 2013). In print, only static 2D projections of the network can be visualized, but the network structure can be further examined with a visualization tool (Bando et al., 2013; Silva et al., 2013). This is important because real system topologies may exhibit very high dimension, hence not suitable to be projected on the plane (Daqing, Kosmidis, Bunde, & Havlin, 2011).

The main topics in a field are associated with communities in the citation networks, which were determined by applying the multilevel community detection method (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). This procedure assigns each paper to a non-overleaping community. It was chosen because it allows for establishing a high modularity for the network, while keeping the computational cost reasonable in comparison with more sophisticated methods such as the optimum modularity (Newman, 2006). By a high modularity we mean that the communities in the network are well distinguishable from each other. It is important to highlight that the multilevel community detection method is stochastic, thus, for each run, a distinct community structure can be attained for the same network. However, as discussed in Blondel et al. (2008), the resulting community partitioning for distinct runs are very similar among themselves and display high correspondence to those obtained by other algorithms or expected from benchmarks.

The relationships among communities were further examined by generating a coarse-grained graph of the network (Rosvall & Bergstrom, 2008), in which each community was replaced by a single community node and its connections. The edges between each pair of community nodes (α, β) were weighted by $W_{\alpha\beta}$ according to the stochastic probability of connections between communities α and β given by:

$$W_{\alpha\beta} = \frac{E_{\alpha\beta}}{|\alpha||\beta|}, \tag{1}$$

where $E_{\alpha\beta}$ is the number of connections among nodes of communities α and β.

Since determining the communities which are most central or peripheral in the science map is an important target, we employed the *accessibility* metric (Amancio, 2015b; Arruda et al., 2014; Travençolo & Costa, 2008; Travençolo, Viana, & Costa, 2009), which is a local node-centered measurement based on the heterogeneity of probabilities of reaching nodes in random walk dynamics. The smaller the accessibility of a node the more peripheral it is. This metric has been successful in separating the topological center and border regions of networks while avoiding the drawbacks of traditional measurements such as betweenness centrality.

Ideally, the communities in the citation network should be labeled with the topics and subtopics of a well-established taxonomy for the scientific field under analysis. However, as already mentioned in Section 1Introduction, there is no simple way to generate such high-level taxonomy automatically. Most authors have therefore resorted to extracting keywords (see Andrade & Valencia, 1998; Carretero-Campos, Bernaola-Galván, Coronado, & Carpena, 2013; Hulth, 2003; Manning & Schütze, 1999 for methods of keyword extraction), for which the majority of the methods make use of large amounts of text. In our case, because we only considered the Abstracts from each paper (representing a node in the network), we had to adapt existing methods. We devised a measurement to quantify the importance of keywords, made with unigrams and bigrams, for each network community. Unigrams and bigrams were extracted for each paper by analyzing its abstract, from which stop-words were removed and the remaining words were lemmatized. This pre-processing step is essential for the analysis because it removes words conveying little semantic content and semantically related words are aliased under the same word if they share the same canonical form (Amancio, 2015a; Amancio, Aluisio, Oliveira, & Costa, 2012; Amancio,

Oliveira, & Costa, 2012d). The *importance index* was designed to quantify the relative frequency of a word appearing inside a community against its frequency on the remainder of the network. First, we count the total number of times $n_\alpha(w)$ a paper presenting a word w appears inside a community α. Next, we calculate the relative in-community frequency, $F_\alpha^{in}(w)$ given by:

$$F_\alpha^{in}(w) = \frac{n_\alpha(w)}{|\alpha|}, \tag{2}$$

where $|\alpha|$ is the number of papers associated with a community α. Analogously, we define a relative out-community frequency:

$$F_\alpha^{out}(w) = \sum_{\gamma \neq \alpha} \frac{n_\gamma(w)}{N - |\alpha|}, \tag{3}$$

which accounts for the total relative frequency considering all communities excluding α, where N is the total number of papers in the network. Then, we define our measurement of importance of keywords, $I(w)$, as the highest difference between the relative in-community and out-community frequencies of a word:

$$I(w) = \max_\alpha [F_\alpha^{in}(w) - F_\alpha^{out}(w)]. \tag{4}$$

The keywords ranked according to the importance index $I(w)$ were used to create trees to simulate the structure of a survey, as shown in Fig. 1. The hierarchy tree (dendrogram) was obtained by a hierarchical agglomerative clustering method (Costa & Cesar, 2009; Duda, Hart, & Stork, 2001), in which we used the average shortest path length, $\langle \ell \rangle_{uv}$, among pairs of keywords (u, v). In this procedure, we first obtained the shortest path lengths ℓ_{ij} between the pairs of papers (i, j) in the citation network. Next, for each keyword pair (u, v) we calculated the average of ℓ_{ij} among pairs of abstracts (A_i, A_j) of papers (i, j), where the keywords u and v were respectively present. This can also be written by the following equation:

$$\langle \ell \rangle_{uv} = \sum_{(u,v) \in (A_i \times A_j)} \frac{\ell_{ij}}{|(u, v) \in (A_i \times A_j)|}. \tag{5}$$

As a consequence, groups of keywords are progressively clustered together according to the average topological distance between them. Therefore, our approach to generating dendrograms incorporates both concepts from complex networks and from text analytics. This was crucial because clustering the keywords using only the Abstracts would not be precise as the amount of text is limited.

Since unigrams and bigrams were ranked according to the same measurement, if a bigram has high $I(w)$, their compounded unigrams are very likely to also feature among the top keywords. To address this problem, we removed the unigrams from the set of keywords that are part of any other bigram in the set. By doing this, we eliminate an immediate layer of redundancy among keywords while also giving priority to more specific keywords (bigrams). For the PC field, we also generated a dendogram using keywords suggested by an expert, in which we omitted generic keywords covering more than 50% of the network (e.g. *photonic crystals* and *fiber*).

The temporal evolution of the fields considered was studied in terms of timelines for the keywords, i.e. how the frequency of each keyword changed over time.

The proposed methodology can be summarized as follows:

1. Obtain the *citation network* among the papers of the corresponding dataset.
2. Obtain the *words*, corresponding to the *n*-grams present on both titles and abstracts of each paper. Here, we considered only unigrams and bigrams for the analysis. Also, we removed stop-words and the remaining words were lemmatized.
3. Apply a *community detection algorithm* to the network, thus obtaining a partitioning of papers. Here, we opted to use a fast multilevel technique (Blondel et al., 2008).
4. Calculate the *in-community frequencies*, $F_\alpha^{in}(w)$, for each word w for all the communities, according to Eq. (2).
5. Calculate the *out-community frequencies*, $F_\alpha^{out}(w)$, according to Eq. (3).
6. Calculate the *importance index*, $I(w)$, of each word w using Eq. (4).
7. Sort the words according to the importance index and select an amount from the top. Here, we selected the first 50 keywords to pair a similar amount of keywords provided by an expert.
8. Apply a hierarchical clustering method to the selected keywords, where the dissimilarity between two keywords corresponds to the average topological distance between papers presenting such words. This procedure results in the dendrogram of keywords.
9. The keywords can also be used to label the communities they belong to.
10. By using network visualization techniques, project the network to a 2D or 3D space and use the communities and the generated labels to obtain a scientific map (Bando et al., 2013; Fruchterman & Reingold, 1991; Silva et al., 2013). In this work we employed the Fruchterman-Reingold algorithm and, for comparison purpose, we also use the VOSViewer (van Eck & Waltman, 2010) visualization tool.

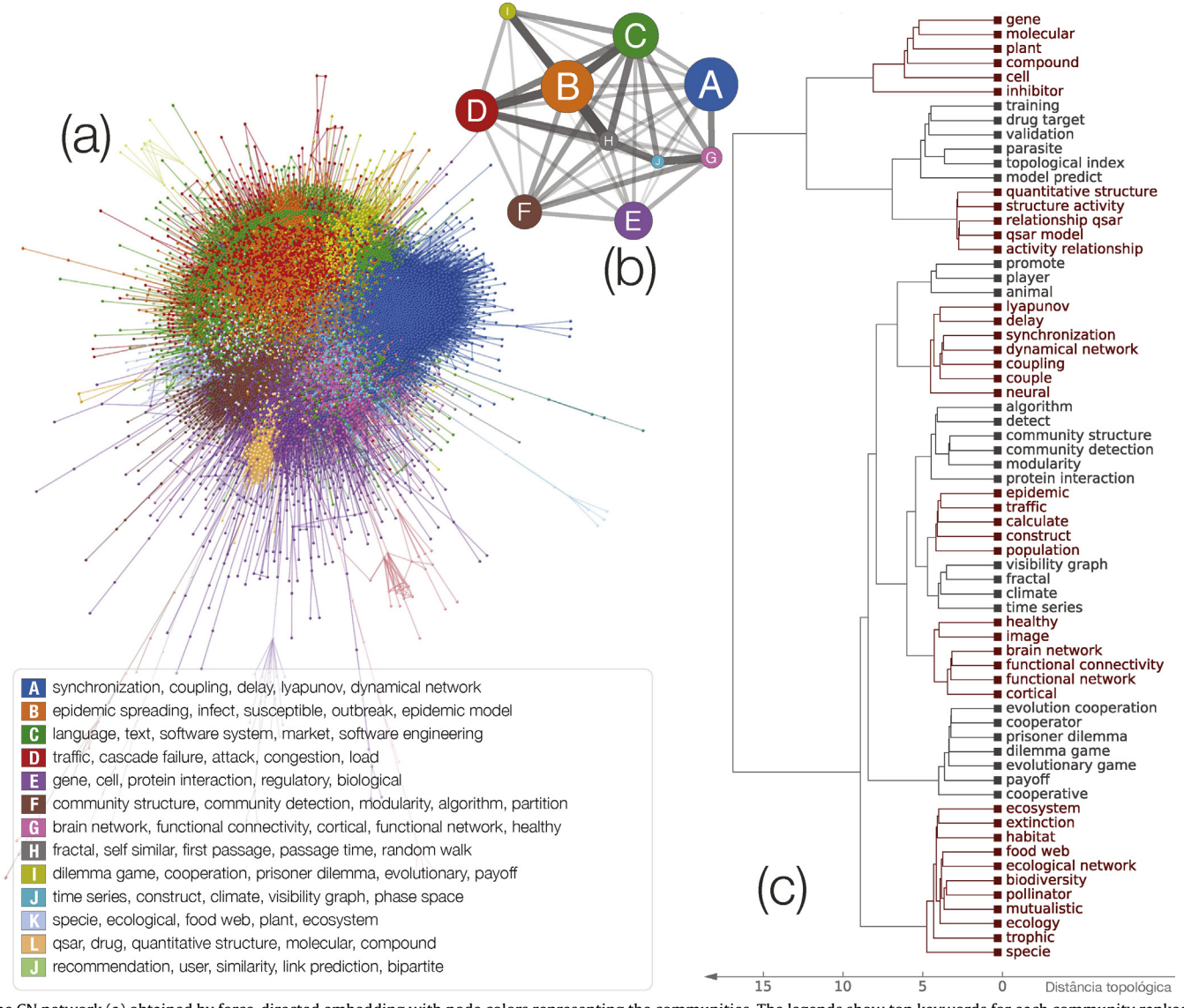

Fig. 2. Projection of the CN network (a) obtained by force-directed embedding with node colors representing the communities. The legends show top keywords for each community ranked according to Eq. (4). The relationships among communities obtained for the CN network are displayed in a coarse-grained diagram (b). The diagram is obtained by collapsing each community in a single node with edges weighted by the fraction of original edges existing against all possible between two communities. Edges are represented by lines with thickness and intensity proportional to their weights. The top 50 keywords for the entire CN network are displayed in a dendrogram (c) built with the hierarchical agglomerative clustering method applied to the topological distance between the keywords.

It should be noted that the techniques employed in each step of our framework can be replaced by similar methods. For instance, one can use other visualization tools and techniques to construct science maps, or one can employ other community detection algorithms. While an extensive combination of techniques and parameters is still needed to uncover benefits and disadvantages of the framework, here we illustrate it by choosing only one set of methods and parameters. These correspond to the most traditional or simple methods required for each step.

4. Results and discussion

We obtained two networks from the dataset, the CN network comprising 11,063 papers with average degree $\langle k_{out}^{CN} \rangle \approx 8.5$, and the PC network encompassing 20,230 papers and presenting $\langle k_{out}^{PC} \rangle \approx 6.6$. Papers published from 1991 to 2013 were included in the networks. The structure of the CN network revealed 22 communities yielding a modularity $q_{CN} \approx 0.53$, while 20 communities were identified with modularity $q_{PC} \approx 0.65$ for the PC network.

4.1. CN network analysis

Fig. 2(a) displays the science map from the CN citation network, where the colors denote the communities associated with the top keywords according to the importance index of Eq. (4). As expected by the high modularity, each module fills distinctive regions of the network topology. The only exception appears to be communities B and D that seem to share the same region, but this is an artifact of the 2D projection. A clear separation is confirmed in the 3D visualization (as shown in Video S1 in Supplementary material). It is interesting that most communities originate from a densely central region of the projection, as can be observed in the figure. This indicates that nodes at the central region are much more interdisciplinary.

The coarse-grained graph of the CN network is shown in Fig. 2(b), which features communities *B*, *C* and *D* strongly connected among themselves. Community *B* (epidemic spreading dynamics) glues together many communities, being at the heart of the network alongside community *H* (fractal, self-similar). This is probably because epidemic dynamics represented by community *B* has a wide variety of applications in network science (Costa et al., 2011). In spite of being the largest community, *A* (synchronization and coupling) only connects strongly to *G* (brain and cortical networks), highlighting the application of synchronization dynamics to modeling neuronal networks. Surprisingly, community *E* (gene regulatory networks, protein interaction, etc) is the lesser connected among the communities. Besides, it presents no remarkable connection preference pattern, i.e. it is uniformly and weakly connected to other communities. This indicates that papers in this community still do not fully benefit from the tools and methodologies provided by network science.

The dendrogram obtained by clustering the top keywords, shown in Fig. 2(c), provided interesting insights. For instance, keywords from the field of ecological applications of complex networks associated with papers containing the words "ecosystem", "food web" and "biodiversity", are closely related among themselves. Although further investigations are needed to explain some counter intuitive exceptions such as the branch containing the keywords "promote", "player" and "animal", on the whole, the relationships between keywords are well described by the dendrogram and appears consistent with what should be expected from an expert in the area.

The analysis was complemented using the *accessibility* metric. The cumulative distribution of accessibility for $h = 3$ taken over all nodes of the CN network is presented in Fig. 3. We chose to calculate accessibility for level $h = 3$ because node-centered measurements taken around the immediate neighborhood of a node (i.e. for $h = 1$ or $h = 2$) may depend on its degree (Costa

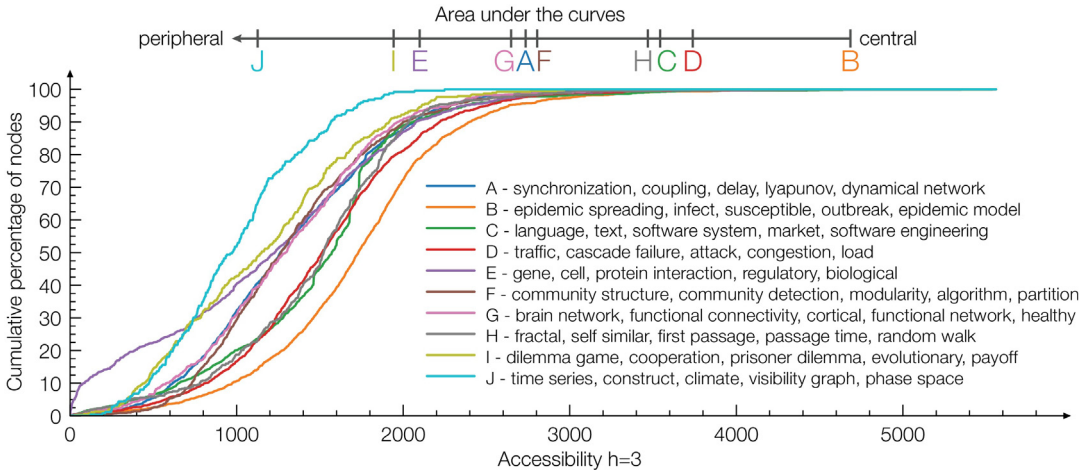

Fig. 3. Curves of cumulative distribution of accessibility obtained for the CN network communities. The curves are presented in color according to the inset. On top of the figure the total area under the curves of each community is shown, which is related to the centrality or peripheral nature of its nodes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

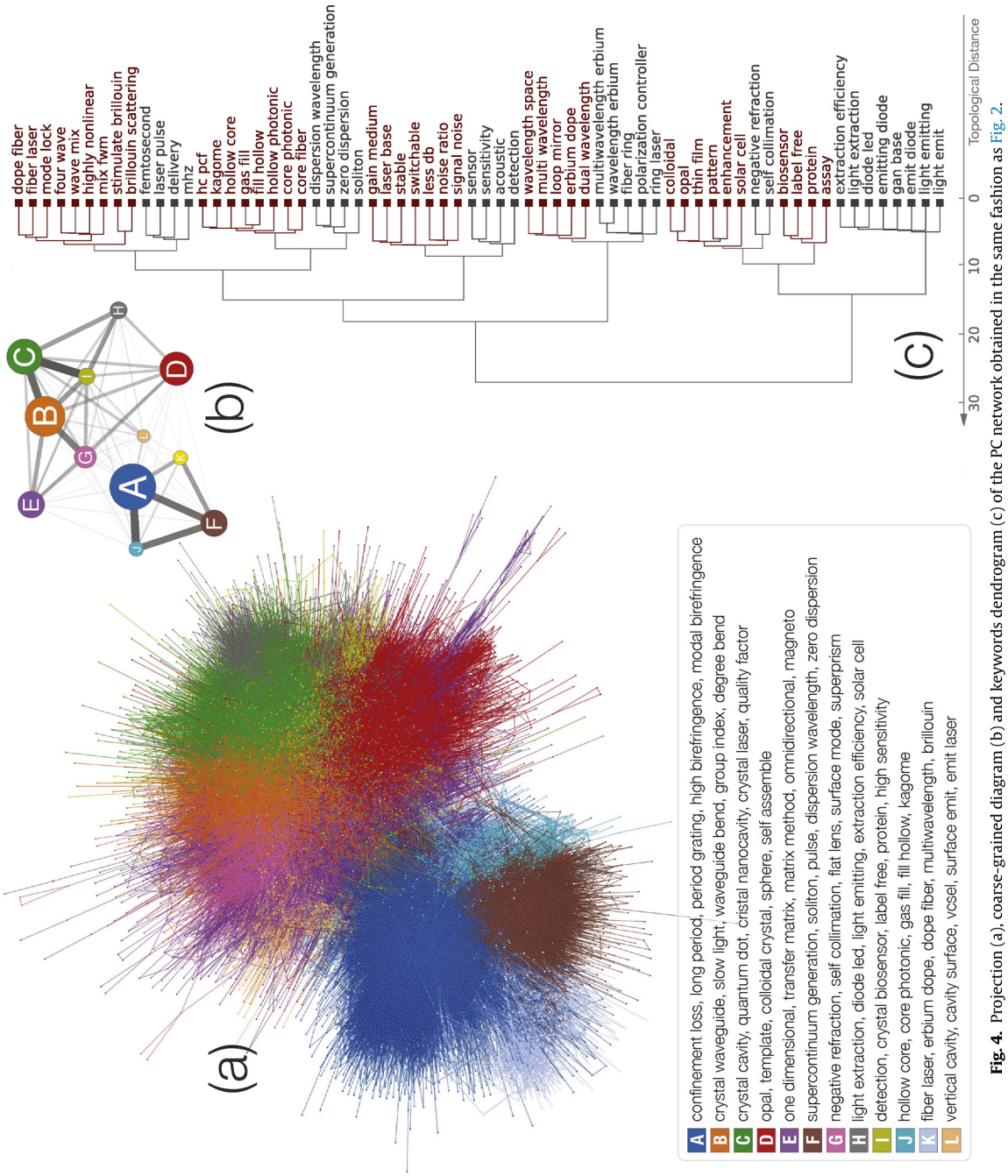

Fig. 4. Projection (a), coarse-grained diagram (b) and keywords dendrogram (c) of the PC network obtained in the same fashion as Fig. 2.

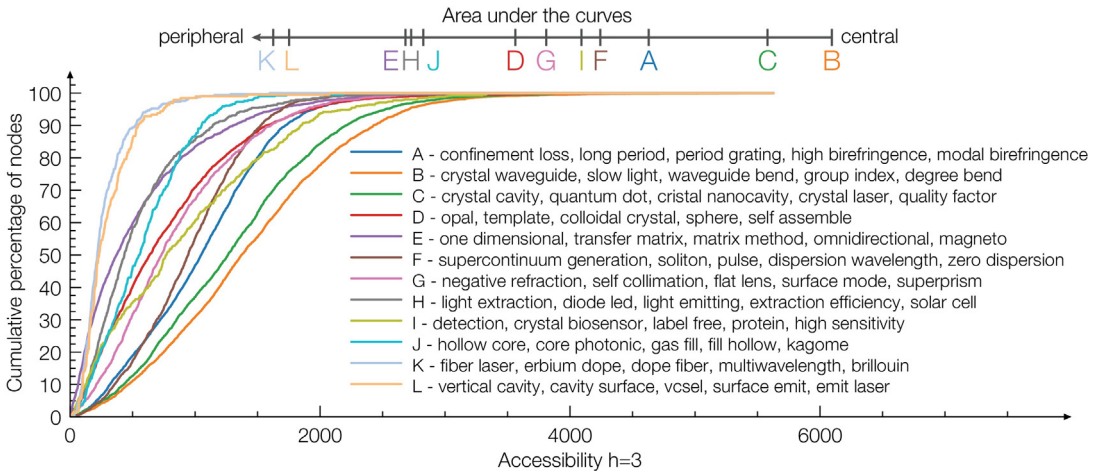

Fig. 5. Cumulative accessibility distribution obtained for the PC network communities.

& Silva, 2006). Also, because the networks are small-world, the measurement may suffer from border effects for large *h*. The data is grouped together by the community membership of nodes, hence each community has a different curve of cumulative accessibility distribution. With the data so presented it is easy to determine the percentage of nodes below or above a certain accessibility threshold. For instance, community *B* possesses only roughly 10% of nodes with accessibility 1000 or lower.

We consider peripheral those communities containing many vertices with low accessibility. The area under the accessibility curves can be used to rank the communities according to their pertinence to the borders of the network. Communities covering a large area under the curves are at the boundaries of the network, as displayed on the top of Fig. 3. Community *J* (time series, climate and visibility graph) is the most peripheral, followed by *I* (game, cooperation and prisoner dilemma) and *E* (protein, gene and cell networks). In particular, community *E* has about 20% of papers with very low accessibility. Communities *G* (brain and cortical networks), *A* (synchronization and coupling) and *F* (community structure and community detection) are close together and present average values of accessibility. The curves for *H* (fractal, self similar and first passage), *C* (language, text and software system) and *D* (traffic, attack, cascade failure) also present similar patterns of accessibility among themselves and are much more at the core of the network than the aforementioned communities.

Corroborating the qualitative results from the analysis of the coarse grained graph, the most central community was *B*. The central core of the network is composed of communities related to techniques of network dynamics such as cascade failure, epidemic spreading dynamics and self-similarity techniques. On the borders are found more specific applications of networks such as cell networks cooperation and time series analysis.

4.2. PC network analysis

The most striking feature of the science map represented by the PC network is its diploid nature, with two very distinct giant communities visualized in Fig. 4(a). From the analysis of keywords associated with these giant communities it is readily noted that they refer to scientists from very distinct areas. The smaller giant community comprises papers from telecommunications, e.g. with keywords deriving from the photonic crystal fiber topic. Indeed, the keywords related to the communities from this giant community are (confinement loss, long period, high birefringence) for *A*, (supercontinuum generation, soliton) for *F*, (fiber laser, erbium dope, dope fiber) for *K* and (porous silicon, silicon photonic, monitor) for *M*. The authors in this giant community are normally engineers exploiting fibers for telecommunications. The larger giant community is made of papers authored by experts in the development of the science of photonic crystals, mostly physicists. The interface between the two giant communities is quite thin, as shown in the figure, thus indicating little scientific interaction across the two enlarged communities.

The interface between the two giant communities is better visualized in the coarse-grained graph in Fig. 4(b), featuring connections from nodes in communities *E* (one dimensional, transfer matrix, matrix method, omnidirectional), *G* (negative refraction, self collimation), *I* (detection, biosensor, label free) and especially *L* (vertical cavity, cavity surface, vcsel, surface emit). Also clear from the coarse-grained graph is the difficulty in establishing which communities are most central or peripheral owing to the diploid nature of the network.

Here is a case where the accessibility metric is most useful. Because it is a local measurement, it avoids the pitfalls of other global centrality measurements when used to characterize networks presenting no well-defined border and central regions. When applied to the PC network, the analysis of cumulative accessibility in Fig. 5 revealed that communities *K* and *L* are those most at the borders, followed by communities *E*, *H* and *J*. Communities *C* and *B* are the most central in the network. Community *A* can also be considered a central community on this smaller giant component. Analogously to what was observed for the CN network, general concepts of the PC field were found in the core of the system, such as papers

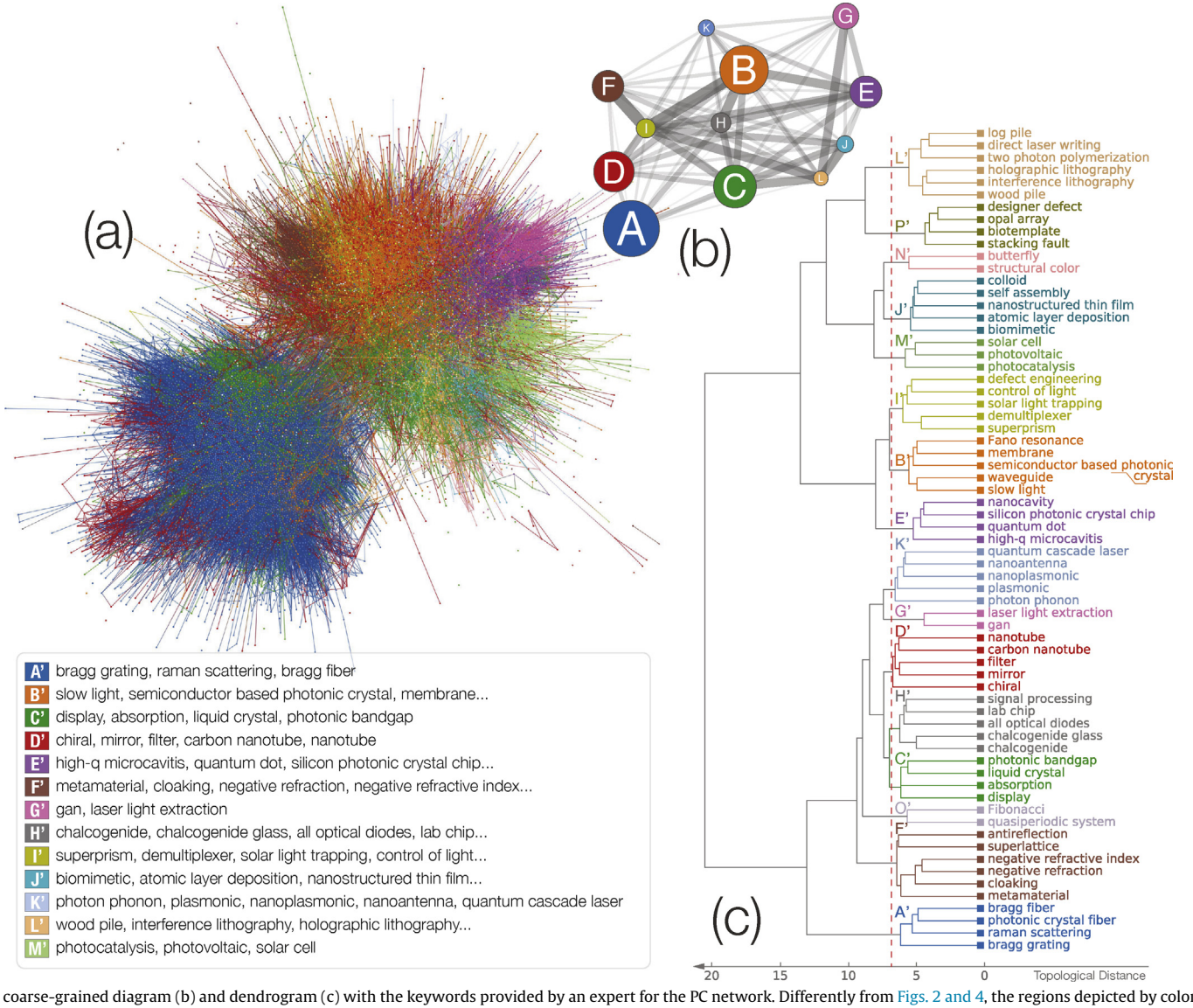

Fig. 6. Projection (a), coarse-grained diagram (b) and dendrogram (c) with the keywords provided by an expert for the PC network. Differently from Figs. 2 and 4, the regions depicted by colors in (a) correspond to the groups obtained after applying a threshold on the dendrogram as indicated by a dashed red line in (c). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

of communities *B* and *C* comprising nodes having keywords “nanocavity”, “quantum dot”, “waveguide”, “slow light”, etc. On the other hand, more specific methodologies and applications are scattered on the borders of the network, such as in papers containing the keywords “fiber laser”, “erbium dope”, “vertical cavity”, “transfer matrix”, “one dimensional”, etc. The taxonomy reached by using the automated keywords for the PC network is consistent with expectation from experts, as indicated in the dendrogram of Fig. 4(c).

We also used a list of 67 keywords, containing up to 4 words each, provided by one of the authors (MB), expert in the PC field. The dendrogram was constructed with the same approach as for the automated keywords in Fig. 6(c). It also provides valuable insights about the area, such as the fact that negative refraction index is closely related to *metamaterials*, which in turn are key concepts for the technology that allows the development of an *invisibility cloak* (Schurig et al., 2006; Soric et al., 2013). Another example concerns the keyword *liquid crystal*, which appears, as expected, close to *photonic bandgap*. A science map of the PC network was obtained using the experts keywords, where partitioning was reached by applying a threshold (as shown by the dashed line and group labels in Fig. 6(c)) to the dendrogram. The nodes were assigned to a community when their corresponding abstracts shared a large number of keywords that define a specific group. A comparison of Figs. 4(a) and 6(a) points to a narrower coverage of nodes for the keywords suggested by the expert for the small giant community associated with the telecommunications area. This was indeed expected because the expert (a physicist) has always worked with topics akin to the large giant community and had less familiarity with the use of photonic crystals in telecommunications. The coarse-grained network shown in Fig. 6(b) bears little resemblance to the one obtained from the community analysis of the network (4), with the groups of the former connecting strongly among themselves. However, a correspondence between some of the network communities and the groups of the experts keywords partitioning can be drawn by observing the communities sharing the same regions of the network (i.e. sharing a similar set of nodes). For instance, community *G* (in Fig. 4(b)) shares the same region as the group *F′* (in Fig. 6(b)), also displaying similar keywords, corresponding to subjects related to negative refraction and cloaking. In the same fashion, communities *C* and *H* share the same region of groups *E′* and *G′*.

To illustrate the possible replacement of methods in one of the steps of our framework, we also imported the network and labeled partitions into the VOSViewer tool (van Eck & Waltman, 2010). This visualization software has been used to construct

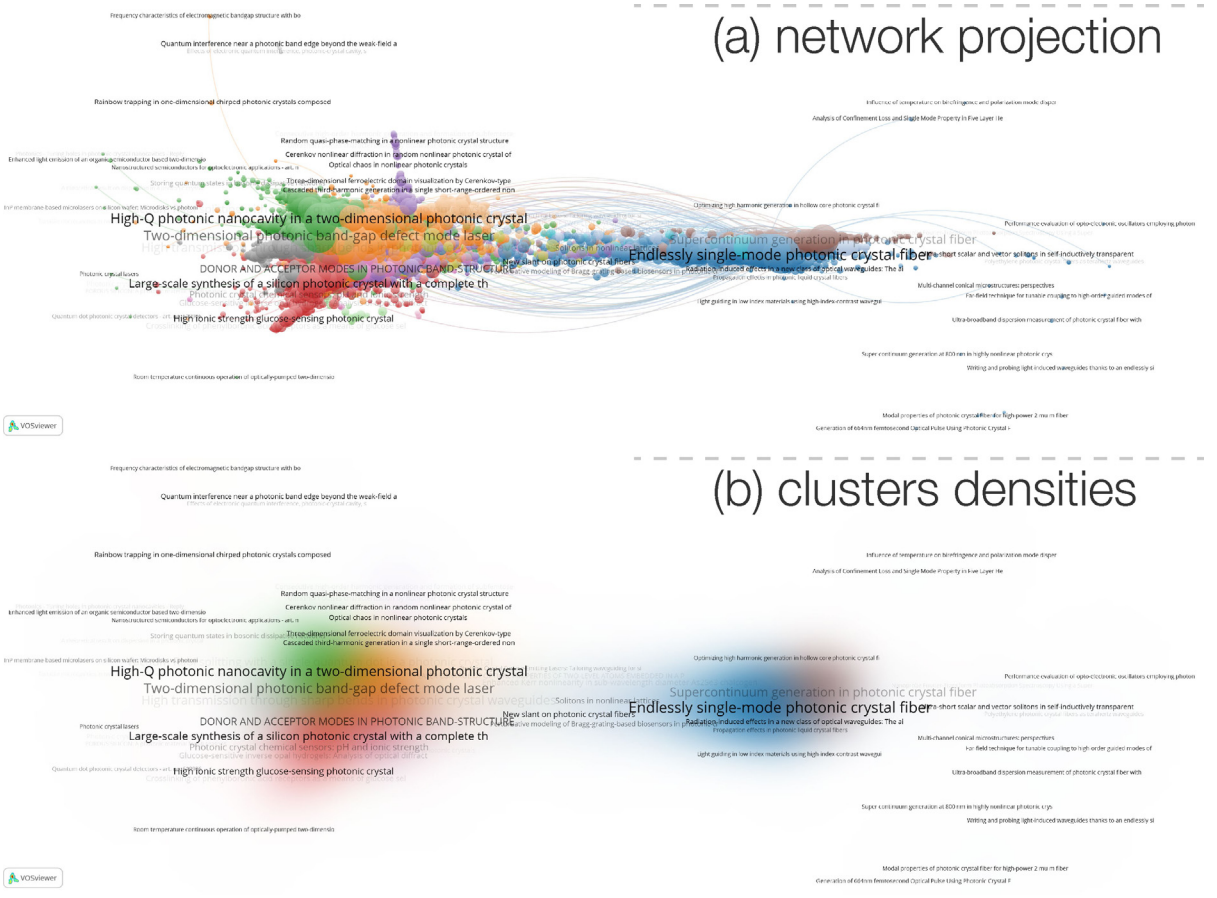

Fig. 7. Visualization of the PC network using the VOSViewer visualization tool (van Eck & Waltman, 2010). The colors represent the same communities displayed in Figure 4. Both the network (a) and densities (b) are shown.

scientific maps from network-based data encompassing a diverse range of disciplines and scientific fields. Fig. 7 displays the projections attained by the software. The existence of the two major groups in the PC network is clearly more accentuated in the VOS Viewer visualization than by using the force-directed method, both in the positions (a) as well as in the density map (b). However, because of the anisotropic nature of the resulting map, some other aspects of the network structure cannot be observed clearly. For instance, it is difficult to tell how interconnected groups *A* and *F* are. In contrast, the isotropic nature of the maps obtained by the force-directed methodology reveals an informative interface between the two groups, which is reflected more clearly by the coarse-grained analysis. Nevertheless, the aims of the visualization techniques are

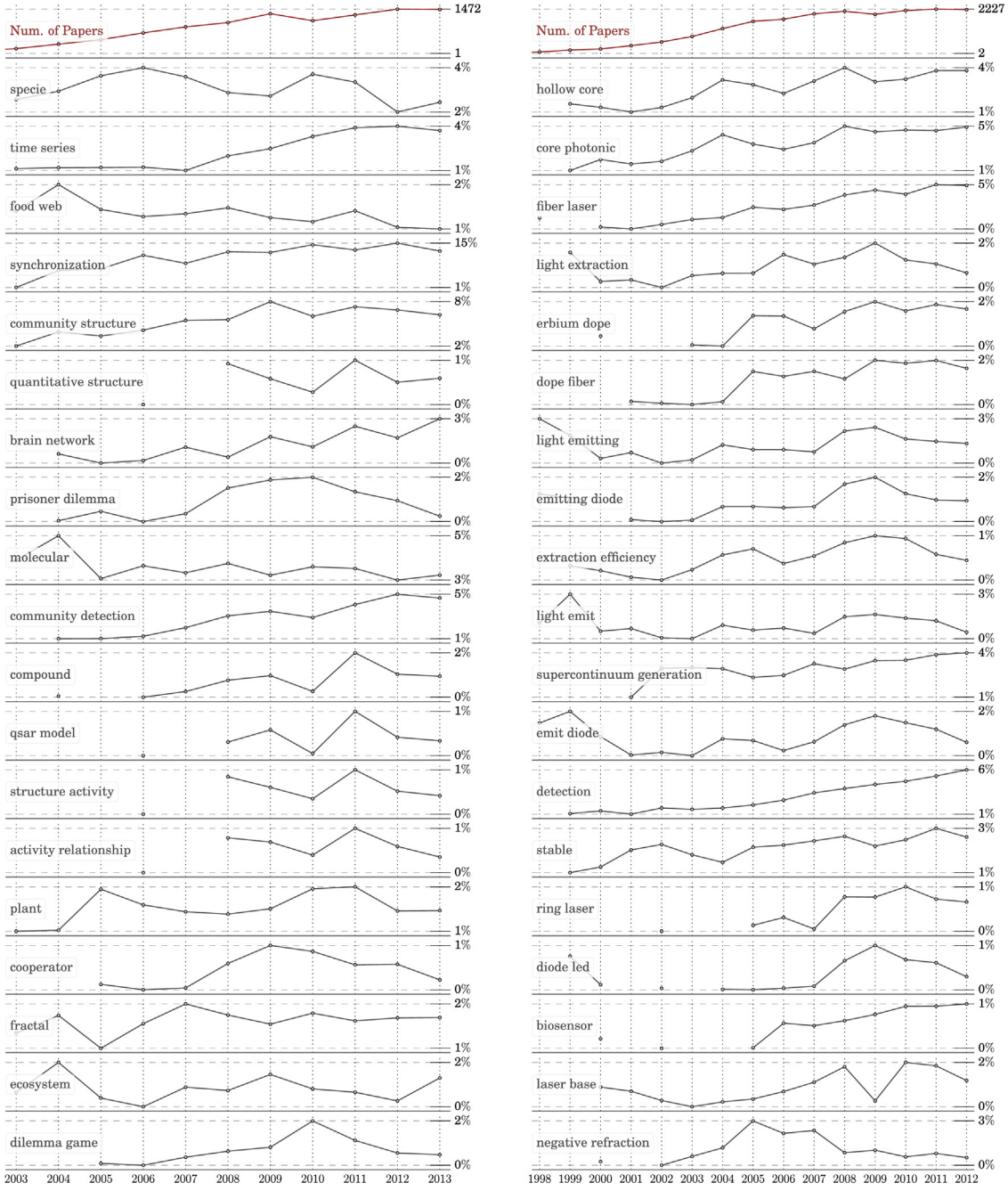

Fig. 8. Normalized frequency of occurrence for each keyword among papers published in the time period considered. The head graphics present the curves corresponding to the number of papers published in the corresponding years.

different and may highlight distinct characteristics of the data. Perhaps the most useful approach is to use as many suitable visualization techniques as possible to draw better conclusions and attain deeper understanding of the datasets and of the analysis.

The temporal evolution of the areas was examined by considering the timeline for the keywords. We counted the number of abstracts which contain the top keywords obtained from the ranking index in Eq. (4). As the number of papers may greatly vary with the years, the frequencies were normalized by the total number of papers published in the same year. The resulting timelines are shown in Fig. 8. Because there are not many papers in the database for the years before 2003 for the CN network, and 1998 for the PC, only the subsequent years were considered.

The timelines confirm the extraordinary growth of both CN and PC areas (as shown on top of Fig. 8), but the growth rate decreased in the last few years. Several areas of CN have been growing: network applications to time series (Donner, Zou, Donges, Marwan, & Kurths, 2010; Lacasa, Luque, Ballesteros, Luque, & Nuño, 2008), synchronization dynamics and analysis (Arenas, Dí az-Guilera, Kurths, Moreno, & Zhou, 2008), community detection (Fortunato, 2010); while other subtopics are shrinking, such as food web and species networks (Dunne, Williams, & Martinez, 2002), cooperation dynamics (Yang, Wang, Wu, Lai, & Wang, 2009) and QSAR model (Santana et al., 2008). In PC field we can also observe distinct growth patterns. The subtopics hollow core photonic, fiber laser, erbium dope fiber, supercontinuum generation, detection, stable and biosensor are still growing on the network, while usage of terms light extraction efficiency, diode led and negative refraction are decreasing.

5. Conclusion and future work

The main goal of this paper was to introduce methods that could be used to automatically construct surveys on a given scientific field. We proposed a methodology to simultaneously analyze contextual information (in terms of papers abstracts) and citation networks, and this was applied to two fields: Complex Networks and Photonic Crystals. Upon identifying communities, it was possible to generate a taxonomy for these fields.

Several patterns could be inferred from the results. For complex networks, for instance, border communities were found to be related to regulatory and protein–protein interaction networks, in addition to subtopics related to climate, time series and visibility graphs. The interpretation is that these subtopics are not fully explored, at the moment, by the many complex networks analysis methods.

The PC network was peculiar in featuring two giant communities, each of which could be identified by analyzing the keywords. As expected, we found that one giant community comprises telecommunication engineers who use photonic crystal fibers in their applications, while the other, larger community is composed mainly of physicists. Surprisingly, not much interaction exists between the two communities, and this piece of information may be valuable to foster collaboration in the future.

The approach proposed here to construct the taxonomy for a survey differs significantly from what exists in the literature. Instead of using only similarities between terms of each abstract, here a citation network was used to provide both the distance among terms and the clustering (derived from the community structure). In addition, a simple text analytics technique was employed to provide the salience of terms according to the obtained community structure.

Here, we did not compare our results to those obtained from traditional text analytics techniques, particularly because the methods address two different classes of problems. Our approach takes into consideration how, in practice, researchers refer to other works in their fields, which may differ significantly from the similarity of terms obtained using only the textual content. The discrepancy between cited works and their contextual similarity has been a recent topic of study, with an in-depth analysis (Amancio et al., 2012c; Ciotti, Bonaventura, Nicosia, Panzarasa, & Latora, 2016). We understand that the organization of the scientific community, i.e., the citation patterns among researchers and papers, must play an important role for constructing a survey in a science field. In this context, our approach is more suitable for this task than methods based solely on text similarity.

We can still compare the technical limitations of the approach presented here and of those based on text analytics. For instance, one of the main disadvantages of topic analysis is the high computational cost involved in estimating the Markov model, which requires several iterations of Gibbs Sampling. This kind of analysis precludes the study of bigrams and higher order n-grams, while our approach can be extended to account for n-grams. In addition, the limitations of such analysis are not yet completely understood (Tang et al., 2014). Other methods such as those based on supervised learning need the input of annotated corpus or sets of golden summaries, which are not commonly available in scientific datasets. We however should point out that our approach is strongly dependent on the chosen network structure. If a co-authorship network among papers was used, instead of the citation network, the results should be interpreted in a different direction and could not be used, for instance, to construct a survey. As for topic analysis, an extensive study of the limitations of our approach is still needed to identify its strengths and disadvantages.

Several extensions of the approach we presented can be performed in future works. For simplicity, we did not consider the direction of the citation networks or the strongly asymmetric nature of the networks. These features could play an important role in the understanding of how distinct fields interact among themselves by citations.

In our methodology we did not take into consideration the importance and redundancy of papers. These limitations may be surpassed by using topological characterization at the level of papers. Future research should also address the problem

of quantifying the interdisciplinarity. It is hoped that the approach inherent in the methods we introduced can be applied to build new tools and assist researchers in understanding their own or new specialty areas.

Acknowledgements

L. da F. Costa thanks CNPq (grant no. 307333/2013-2) and FAPESP-MCT/CNPq/PRONEX (grant no. 11/50761-2) for support. F.N. Silva acknowledges CAPES and FAPESP (grant no. 15/08003-4). D.R. Amancio thanks FAPESP (grant no. 14/20830-0). O.N. Oliveira Jr acknowledges FAPESP and CNPq. M. Bardosova acknowledges Science Foundation Ireland for support.

Author contributions

Conceived and designed the analysis: Filipi N. Silva, Diego R. Amancio, Luciano da F. Costa, and Osvaldo N. Oliveira Jr.
Collected the data: Filipi N. Silva and Maria Bardosova.
Contributed data or analysis tools: Filipi N. Silva and Diego R. Amancio.
Performed the analysis: Filipi N. Silva, Diego R. Amancio, Maria Bardosova, Luciano da F. Costa, and Osvaldo N. Oliveira Jr.
Wrote the paper: Filipi N. Silva, Diego R. Amancio, and Osvaldo N. Oliveira Jr.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.joi.2016.03.008.

References

Amancio, D. R. (2015a]). Comparing the topological properties of real and artificially generated scientific manuscripts. *Scientometrics*, *105*(3), 1763–1779.
Amancio, D. R. (2015b]). A complex network approach to stylometry. *PLoS ONE*, *10*(8), 1–21.
Amancio, D. R., Aluisio, S. M., Oliveira, O. N., Jr., & Costa, L. da F. (2012). Complex networks analysis of language complexity. *EPL (Europhysics Letters)*, *5*(10), 58002.
Amancio, D. R., Nunes, M. G. V., Oliveira, O. N., Jr., & Costa, L. da F. (2012b]). Extractive summarization using complex networks and syntactic dependency. *Physica A: Statistical Mechanics and Its Applications*, *391*(4), 1855–1864.
Amancio, D. R., Nunes, M. G. V., Oliveira, O. N., Jr., & Costa, L. da F. (2012c]). Using complex networks concepts to assess approaches for citations in scientific papers. *Scientometrics*, *91*(3), 827–842.
Amancio, D. R., Oliveira, O. N., Jr., & Costa, L. da F. (2012d]). Identification of literary movements using complex networks to represent texts. *New Journal of Physics*, *14*(4), 043029.
Amancio, D. R., Oliveira, O. N., Jr., & Costa, L. da F. (2012e]). On the use of topological features and hierarchical characterization for disambiguating names in collaborative networks. *EPL (Europhysics Letters)*, *99*(4), 48002.
Andrade, M. A., & Valencia, A. (1998). Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families. *Bioinformatics*, *14*(7), 600–607.
Arenas, A., Dí az-Guilera, A., Kurths, J., Moreno, Y., & Zhou, C. (2008). Synchronization in complex networks. *Physics Reports*, *469*(3), 93–153.
Arruda, G. F., Barbieri, A. L., Rodrí guez, P. M., Rodrigues, F. A., Moreno, Y., & Costa, L. da F. (2014). Role of centrality for the identification of influential spreaders in complex networks. *Physical Review E*, *90*(3), 032812.
Bando, S. Y., Silva, F. N., Costa, L. da F., Silva, A. V., Pimentel-Silva, L. R., et al. (2013). Complex network analysis of ca3 transcriptome reveals pathogenic and compensatory pathways in refractory temporal lobe epilepsy. *PLoS ONE*, *8*(11), e79913.
Barabasi, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.
Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. *Science*, *323*(5919), 1297–1298.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.
Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.
Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., et al. (2012). Design and update of a classification system: The UCSD map of science. *PLoS ONE*, *7*(7)
Börner, K., & Scharnhorst, A. (2009). Visual conceptualizations and models of science. *Journal of Informetrics*, *3*(3), 161–172.
Boyack, K., Klavans, R., & Borner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351–374.
Boyack, K. W., & Klavans, R. (2014). Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology*, *65*(4), 670–685.
Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., et al. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS ONE*, *6*(3), e18029.
Boyack, K. W., Wylie, B. N., & Davidson, G. S. (2002). Domain visualization using VxInsight (r) for science and technology management. *Journal of the Association for Information Science and Technology*, *53*(9), 764–774.
Carretero-Campos, C., Bernaola-Galván, P., Coronado, A., & Carpena, P. (2013). Improving statistical keyword detection in short texts: Entropic and clustering approaches. *Physica A: Statistical Mechanics and Its Applications*, *392*(6), 1481–1492.
Chen, C. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl. 1), 5303–5310.
Chen, C., & Hicks, D. (2004). Tracing knowledge diffusion. *Scientometrics*, *59*(2), 199–211.
Chen, T. T. (2012). The development and empirical study of a literature review aiding system. *Scientometrics*, *92*(1), 105–116.
Ciotti, V., Bonaventura, M., Nicosia, V., Panzarasa, P., & Latora, V. (2016). Homophily and missing links in citation networks. *EPJ Data Science*, *5*(1).
Costa, L. da F., & Cesar, R. M., Jr. (2009). *Shape Classification and analysis: theory and practice* (2nd ed.). *Vol. 10 of Image processing series* CRC Press.
Costa, L. da F., Oliveira, O. N., Jr., Travieso, G., Rodrigues, F. A., Villas Boas, P. R., et al. (2011). Analyzing and modeling real-world phenomena with complex networks: A survey of applications. *Advances in Physics*, *60*(3), 329–412.
Costa, L. da F., & Silva, F. N. (2006). Hierarchical characterization of complex networks. *Journal of Statistical Physics*, *125*, 845–876.
Craddock, T., Harwood, C. R., Hallinan, J., & Wipat, A. (2008). e-Science: Relieving bottlenecks in large-scale genome analyses. *Nature Reviews Microbiology*, *6*(12), 948–954.
Daqing, L., Kosmidis, K., Bunde, A., & Havlin, S. (2011). Dimension of spatially embedded networks. *Nature Physics*, *7*(6), 481–484.
Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009). Pagerank for ranking authors in co-citation networks. *Journal of the Association for Information Science and Technology*, *60*(11), 2229–2243.

Donner, R. V., Zou, Y., Donges, J. F., Marwan, N., & Kurths, J. (2010). Recurrence networks – A novel paradigm for nonlinear time series analysis. *New Journal of Physics*, *12*(3), 033025.
Donovan, S. (2008). Big data: Teaching must evolve to keep up with advances. *Nature*, *455*(7212), 461.
Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. John Wiley and Sons.
Dunne, J. A., Williams, R. J., & Martinez, N. D. (2002). Food-web structure and network theory: The role of connectance and size. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(20), 12917–12922.
Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479.
Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3–5), 75–174.
Fruchterman, T., & Reingold, E. (1991). Graph drawing by force-directed placement. *Software-Practice & Experience*, *21*, 1129–1164.
Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl. 1), 5228–5235.
Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In M. Collins, & M. Steedman (Eds.), *Proceedings of the 2003 conference on empirical methods in natural language processing* (pp. 216–223).
Jenssen, T.-K., Lægreid, A., Komorowski, J., & Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, *28*(1), 21–28.
Kedzie, C., McKeown, K., & Diaz, F. (2015). Predicting salient updates for disaster summarization. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing of the Asian Federation of natural language processing, ACL 2015, Vol. 1: Long Papers* (pp. 1608–1617).
Lacasa, L., Luque, B., Ballesteros, F., Luque, J., & Nuño, J. C. (2008). From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(13), 4972–4975.
Laporte, G. (1992). The vehicle routing problem: An overview of exact and approximate algorithms. *European Journal of Operational Research*, *59*(3), 345–358.
Lariviere, V., Haustein, S., & Börner, K. (2015). Long-distance interdisciplinarity leads to higher scientific impact. *PLoS ONE*, *10*(3).
Lawrence, S., & Giles, C. L. (1998). Searching the world wide web. *Science*, *280*(5360), 98–100.
Leicht, E. A., Clarkson, G., Shedden, K., & Newman, M. E. J. (2007). Large-scale structure of time evolving citation networks. *The European Physical Journal B*, *59*(1), 75–83.
Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27) (pp. 2177–2185). Curran Associates, Inc.
Leydesdorff, L., de Moya-Anegón, F., & de Nooy, W. (2015). Aggregated journal–journal citation relations in scopus and web of science matched and compared in terms of networks, maps, and interactive overlays. *Journal of the Association for Information Science and Technology*.
Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the isi subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348–362.
Leydesdorff, L., Rafols, I., & Chen, C. (2013). Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal–journal citations. *Journal of the American Society for Information Science and Technology*, *64*(12), 2573–2586.
Li, J., Zhou, J., Luo, X., & Yang, Z. (2012). Chinese lexical networks: The structure, function and formation. *Physica A: Statistical Mechanics and Its Applications*, *391*(21), 5254–5263.
Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA, USA: MIT Press.
McKeown, K., Daume, H., Chaturvedi, S., Paparrizos, J., Thadani, K., Barrio, P., et al. (2016). Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology*.
Menczer, F. (2004). Correlated topologies in citation networks and the web. *The European Physical Journal B*, *38*(2), 211–221.
Mullen, E. K., Daley, M., Backx, A. G., & Thompson, G. J. (2014). Gene co-citation networks associated with worker sterility in honey bees. *BMC Systems Biology*, *8*(38).
Nenkova, A., & McKeown, K. (2012). A survey of text summarization techniques. In *Mining text data*. pp. 43–76. Springer.
Newman, M. (2010). *Networks: An introduction*. New York, NY, USA: Oxford University Press, Inc.
Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(23), 8577–8582.
Newman, M. E. J., & Clauset, A. (2015). *Structure and inference in annotated networks*. arXiv:1507.04001
Nichols, J., Mahmud, J., & Drews, C. (2012). *Summarizing sporting events using twitter*. ACM.
Nykl, M., Campr, M., & Ježek, K. (2015). Author ranking based on personalized pagerank. *Journal of Informetrics*, *9*(4), 777–799.
Porter, A., & Rafols, I. (2009). Is science becoming more interdisciplinary? measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719–745.
Radev, D. R., & McKeown, K. R. (1998). Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, *24*(3), 470–500, 0891-2017.
Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(4), 1118–1123.
Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.
Santana, L., González-Dí az, H., Quezada, E., Uriarte, E., Yáñez, M., Viña, D., et al. (2008). Quantitative structure–activity relationship and complex network approach to monoamine oxidase a and b inhibitors. *Journal of Medicinal Chemistry*, *51*(21), 6740–6751.
Schurig, D., Mock, J. J., Justice, B. J., Cummer, S. A., Pendry, J. B., Starr, A. F., et al. (2006). Metamaterial electromagnetic cloak at microwave frequencies. *Science*, *314*(5801), 977–980.
Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47.
Shen, C., Liu, F., Weng, F., & Li, T. (2013). A participant-based approach for event summarization using twitter streams. In *Proceedings of NAACL* (pp. 1152–1162).
Silva, F. N., & Costa, L. da F. (2011). *Network 3D*. http://cyvision.ifsc.usp.br/Cyvision/?page=SOFTWARE&subpage=NETWORKS3D
Silva, F. N., Rodrigues, F. A., Oliveira, O. N., Jr., & Costa, L. da F. (2013). Quantifying the interdisciplinarity of scientific journals and fields. *Journal of Informetrics*, *7*(2), 469–477.
Silva, F. N., Travençolo, B. A. N., Viana, M. P., & Costa, L. da F. (2010). Identifying the borders of mathematical knowledge. *Journal of Physics A: Mathematical and Theoretical*, *43*(32), 325202.
Silva, F. N., Viana, M. P., Travençolo, B. A. N., & Costa, L. da F. (2011). Investigating relationships within and between category networks in wikipedia. *Journal of Informetrics*, *5*(3), 431–438.
Silva, T. C., & Amancio, D. R. (2012). Word sense disambiguation via high order of learning in complex networks. *EPL (Europhysics Letters)*, *98*(5), 58001.
Skupin, A., Biberstine, J. R., & Börner, K. (2013). Visualizing the topical structure of the medical sciences: A self-organizing map approach. *PLoS ONE*, *8*(3).
Soric, J. C., Chen, P. Y., Kerkhoff, A., Rainwater, D., Melin, K., & Alù, A. (2013). Demonstration of an ultralow profile cloak for scattering suppression of a finite-length rod in free space. *New Journal of Physics*, *15*(3), 033037.
Stone, R., & Lavine, M. (2014). The social life of robots. *Science*, *346*(6206), 178–179.
Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. In T. Jebara, & E. P. Xing (Eds.), *Proceedings of the 31st International Conference on Machine Learning (ICML-14). JMLR workshop and conference proceedings* (pp. 190–198).

Travençolo, B. A. N., Viana, M. P., & Costa, L. da F. (2009). Border detection in complex networks. *New Journal of Physics*, *11*, 063019.
Travençolo, B., & Costa, L. da F. (2008). Accessibility in complex networks. *Physics Letters A*, *373*(1), 89–95.
Üsdiken, B., & Pasadeos, Y. (1995). Organizational analysis in north America and Europe: A comparison of co-citation networks. *Organization Studies*, *16*(3), 503–526.
van Eck, N. J., & Waltman, L. (2010). Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.
van Eck, N. J., & Waltman, L. (2014). Citnetexplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, *8*(4), 802–823.
Viana, M. P., Travencolo, B. A. N., Tanck, E., & Costa, L. da F. (2010). Characterizing topological and dynamical properties of complex networks without border effects. *Physica A: Statistical Mechanics and Its Applications*, *389*(8), 1771–1778.
Waaijer, C. J. F., van Bochove, C. A., & van Eck, N. J. (2011). On the map: Nature and science editorials. *Scientometrics*, *86*(1), 99–112.
Yan, X., Zhai, L., & Fan, W. (2013). C-index: A weighted network node centrality measure for collaboration competence. *Journal of Informetrics*, *7*(1), 223–239.
Yang, H.-X., Wang, W.-X., Wu, Z.-X., Lai, Y.-C., & Wang, B.-H. (2009). Diversity-optimized cooperation on complex networks. *Physics Review E*, *79*, 056107.
Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys*, *35*(4), 399–458.
Zhou, J., Zeng, A., Fan, Y., & Di, Z. (2015). Ranking scientific publications with similarity-preferential mechanism. *Scientometrics*, *106*(2), 805–816.