Chaos, Solitons and Fractals 142 (2021) 110509

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

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

Enriched line graph: A new structure for searching language collocations

Ángeles Criado-Alonso[a,b,*], Elena Battaner-Moro[a,c], David Aleja[d,f], Miguel Romance[d,e,f], Regino Criado[d,e,f]

[a] Grupo de Investigación LIyNMEDIA, Universidad Rey Juan Carlos c/Tulipán s/n, Móstoles 28933 Madrid, Spain
[b] Departamento de Economa Financiera y Contabilidad e Idioma Moderno, Spain
[c] Departamento de CC. de la Educación, Lenguaje, Cultura y Artes, Ciencias Histórico-Jurídicas y Humanísticas y Lenguas Modernas, Spain
[d] Departamento de Matemática Aplicada, Ciencia e Ingeniería de los Materiales y Tecnología Electrónica, ESCET Universidad Rey Juan Carlos, C/Tulipán s/n, Móstoles 28933 Madrid, Spain
[e] Center for Computational Simulation, Universidad Politécnica de Madrid, Pozuelo de Alarcón 28223 Madrid, Spain
[f] Data, Complex Networks and Cybersecurity Sciences Technological Institute, Univ. Rey Juan Carlos, Madrid 28028, Spain

ARTICLE INFO

Article history:
Received 3 November 2020
Accepted 22 November 2020
Available online 28 November 2020

Keywords:
Enriched line graph
Multilayer networks
PageRank
Interaction of words
Language collocations

ABSTRACT

The specific terminology of a specialty language comes, essentially, from specific uses of already existing words and/or from specific combinations of words so called “collocations”. In this work we introduce a new mathematical structure (enriched line graph) and a new methodology to extract properties and characteristics of a type of multilayer linguistic networks associated with these types of languages. Specifically, this work is focused on the description of a methodology based on a variant of the PageRank algorithm to locate the linguistic collocations and on defining a new structure (enriched line graph) that can be interpreted as a certain type of “interpolation” between the original graph and its associated line graph, showing new results, properties and applications of this concept, and, in particular, certain characteristics of the specialty language produced by the scientific community of complex networks.

1. Introduction

In the last two decades, the emergence of different types and models of networks in modeling the interactions between different parts of a complex system is a popular and well-known fact in virtually all areas of knowledge that have benefited from the advances in the study of complex networks [2,3,6,7,17,27,35,37,46,52]. Complex networks have emerged as an indispensable tool for simulating the interactions and relationships between the components of a complex system in domains as diverse as technology, biology and human social organization. Analyzing a particular system, identifying a complex network related to it and exploring the properties of that network to learn about the system under analysis is a strategy that has provided a great number of very significant results in many real applications [2,6,7,13,23,46,52]. The latest advances in modern linguistics are based on the treatment of a language as a system or a complex network, to which the tools, measures and procedures of this branch of science can be applied in order to obtain a new, efficient and effective approach to the study of language [9,14,24,28,29,38,40,42,44,51]. That is why the analysis of linguistic theories supported by the study of specialized corpora and the new vision provided by complex networks allows us to obtain certain stylistic and typological characteristics and some intrinsic properties of languages. In order to do that, we perform a computerized treatment on a linguistic corpus, i.e., a collection of texts collected electronically according to a set of specific criteria used as a representative sample of a language or subset of that language [10,41]. To carry out this research, a linguistic corpus is set up: it is composed by 86 extended abstracts and papers (volumes 1–6 of the International Journal of Complex Systems in Science (IJCSS), published between April 2011 and November 2016 (http://www.ij-css.org)), giving a total amount of 147,637 words and 25,210 sentences analyzed in this study. This complex network (see Figs. 1 and 2) is been used to design a help prototype of aid tool for specialized translations of this scientific area. The unit of analysis considered in this paper is the sentence, that is, the words enclosed between two points [32]. Also, it is important to note that commas and other punctuation marks within the

* Corresponding author at: Grupo de Investigación LIyNMEDIA, Universidad Rey Juan Carlos c/Tulipán s/n, 28933 Móstoles Madrid, Spain.
E-mail addresses: angeles.criado@urjc.es (Á. Criado-Alonso), elena.battaner@urjc.es (E. Battaner-Moro), david.aleja@urjc.es (D. Aleja), miguel.romance@urjc.es (M. Romance), regino.criado@urjc.es (R. Criado).

https://doi.org/10.1016/j.chaos.2020.110509
0960-0779/

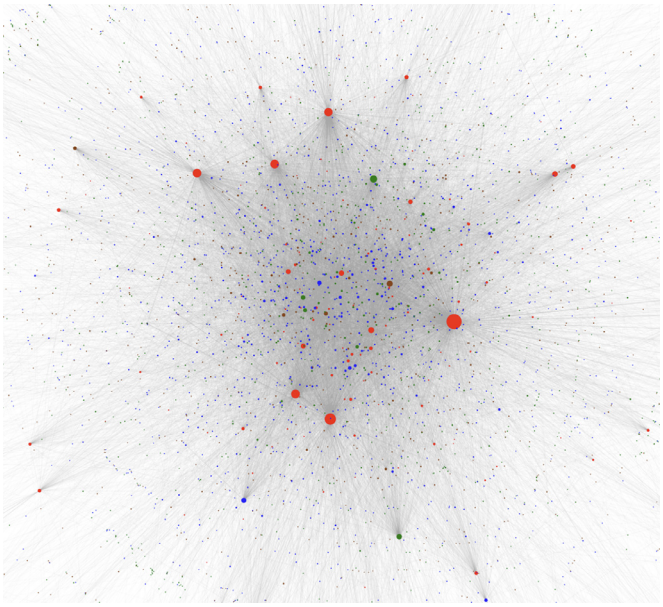

Fig. 1. Linguistic Network composed by 147,637 words of which 1,487 are lexical words. The size of nodes is proportional to their out-degree and its colour indicates the layer it belongs to.

sentence have been removed for the analysis done. The study of frequent words or the lexical behavior of a set of words, which is usually the aim of many works of this type, has been, in our case, only the first step. The co-presence or co-occurrence of other words that construct new units of meaning is necessary to deepen in linguistic issues.

It should be highlighted that in this work the methodology of complex networks is used for combinatorial analysis of words (syntagmatic approach, from the Greek "$\sigma\iota\upsilon\tau\alpha\gamma\mu\alpha$", syntagma: "group assembled") since, according to Firth [30], "the syntagmatic relationship at the lexical level is called *collocation*".

It is important to point out that the type of analysis presented in this paper, the syntagmatic approach, is different from the paradigmatic approach common to other works in computational linguistics. This kind of approach consists of studying the relationships between words within the paradigm, understood as the set of words that can appear in that place of a linguistic structure. For example, if we have an (adjective + noun) construction with a syntactic function (i.e. the subject of a sentence), the first element paradigm is the set of words that can appear in that position (i.e. the adjectives) and the second element paradigm is the set of words that can appear in that position (i.e. the nouns). The paradigmatic relationship is based on the value by opposition that the linguistic elements possess.

On the other hand, the syntagmatic relationship is a study based on the relationships of words with each other within the linguistic structure. At the lexical level, on which our study focuses, we can find syntagmatic relationships of this type in the so-called *collocations* [49], that is, words that appear together in language, so that the meaning of the collocation is not the result of the sum of two meanings ("black mail"/"black trousers"), but that they acquire a new meaning that is only possible when these two words appear together (and they are not interchangeable: there is no "white mail" but there are "white trousers"). This kind of syntagmatic relationship is common in certain verbs (phrasal verbs): examples such as "to work out", "to turn down", "to turn up" do not acquire their meaning from the sum of their individual meanings, but have one of their own just because they go together (syntagmatic relationship) or only when they go together.

The analysis presented here allows us to support the initial hypothesis (the syntagmatic study) with a methodology and analysis that has made it possible to locate and specify new units of meaning specific to this specialty language (i.e., collocations) and to study their behaviors and their characteristic weight in the corpus.

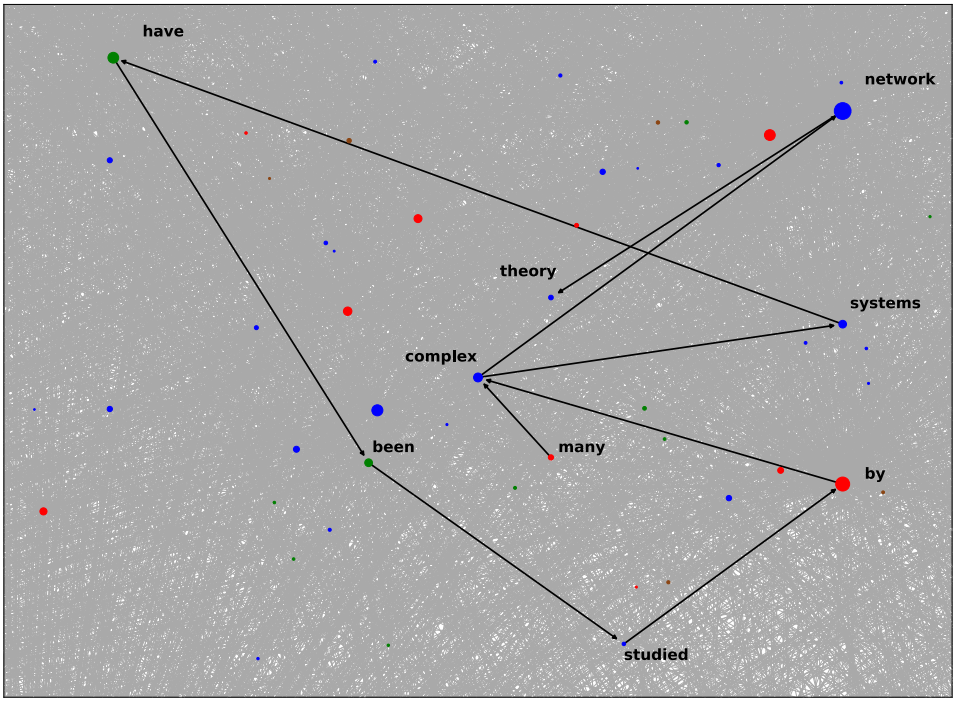

Fig. 2. Zoom of the multilayer linguistic complex network studied. The color of each node indicates the layer it belongs to.

The latest advances in the study of complex systems that have made it possible to move from the analysis and study of the isolated properties of a system to a more realistic modeling in which multiple phenomena of a different nature interact and coexist, lead and motivate the use of multilayer networks for the analysis of the structure and characterization of speciality languages. It is important to mention that although the model provided by multilayer networks has recently appeared in the literature, this model has already been used as a suitable structure for the structural analysis of language systems [42,44]. In any case, the application of theory and computational tools of complex networks to the study of languages is not a new scientific trend [9,14,28,29,38–40,43,50,54]. Multilayer networks constitute a unified framework that allows the modeling of the structural properties of specialty languages by exploring the interaction between linguistic units and the formation of associations of these units, constituting new structures that will be part of the specialty language terminology.

So, in this work, to analyze the collocations as parts of the language terminology, a multilayer network and the concept of line graph [16,53] have been considered using the same type of structure as in [21].

The structure of the paper is as follows. After this introduction, in Section 2 a linguistic multilayer network is introduced and some remarkable results related to this new structure are obtained. Section 3 is devoted to introduce the mathematical concepts of enriched line graph and semi-line graph of a weighted network and some results and relationships between this new concepts and the concept of line graph are presented. In Section 4 both new concepts are extended to multilayer networks with the aim of obtaining some applications and results related to the multilayer linguistic network model. In Section 5 we present some numerical experiments and applications to illustrate the results of the previous sections. Section 6 is devoted to present some characteristics of the specialty language generated by the complex network science community. Finally in Section 7 we present some conclusions of this work.

2. The multilayer linguistic network model

The linguistic corpus under our study consists of texts written in English. The English language has four major word classes: nouns, adjectives, verbs, and adverbs. Other word classes are prepositions, conjunctions, determiners, interjections, or pronouns [34]. On this basis we have established the layers with the aim of studying the lexical behavior of certain words in the specialty language. In this regard, a methodology close to supervised machine learning has been used, since in order to discriminate between the terms of the corpus used and to assign them to one or another layer, a completely lexical linguistic decision was made according to the criteria of several experts in this specialty language, so that all the terms of the corpus have been labeled and distributed in the different layers according to their morphological and lexical properties in successive approximations until the current distribution in layers has been reached. This methodology of analysis has led us to consider four layers (lexical layer, verb layer, linking layer and remaining words layer) that have already been used in the seminal work [21] and which can be exportable to the study for the analysis of any other specialty language. This multilayer network approach allows us to characterize different types of interactions between nodes, which is not possible with the traditional monolayer network approach (e.g., the type of interaction between nodes in the verb layer versus the interaction of a node in that layer with one in the lexical layer). Moreover, this structure allows us to contemplate the presence of the same term present in two different layers, according to the role played (multiplex network). As an example, “model” is a node of the verbal layer and also of the lexical layer, and the same happens with “cluster”, a term that if it appears followed by ‘together’ would be in the verbal layer, and otherwise in the lexical layer. The recognition of a network formed by four layers is necessary in order to be able to discriminate a syntactic relationship (e.g., a noun followed by a verb, i.e., subject + predicate, or the presence of prepositions) from a semantic relationship (combinations of nouns and adjectives) which, in fact, is the purpose of our research. This division into four layers has been a linguistic (word categories) and technical (specialty vocabulary in complex networks) decision. In this regard, it is easy to understand that, for example, in the context of the specialty language of complex networks, the meaning of a combination such as “complex network” is not only the sum of the meaning of “complex” + “network” (taken as single words), but can be conceived as a precise unit of meaning independent of its appearance individually or together with other words (v.g. “railway network”). Hence, we take “complex network” as a collocation and consequently it becomes a lexical unit within the specialty language of complex networks.

The behavior of the nodes is essentially different depending on the layer they belong to: we analyze the intra-layer connections, for example in the case of the lexical layer and the collocations (elements of the same layer that go together) and also, the inter-layer connections which are the basic grammatical relations in a sentence. In any case, the model described is a particular case of a multilayer network [7].

It is the interaction between layers that facilitates the description of the formation of specialty verbs. Moreover, taking into account that the syntagma is the basic syntactic unit, the model allows us to describe the conversion of these units into new units according to their symptomatic meaning (constructions of more than one word from two or more that go together in a sentence).

If we select the words and layer them, the semantic information they give us is completely different. In particular, these tables justify why we use the PageRank algorithm and not simply the frequencies. Frequency tells us nothing about the relationships between words, while PageRank does. For example, “in” and “and” change their position depending on whether one considers the frequency or the PageRank. As the frequency of the lexical words increases, all the nodes in the lexical layer should rise linearly, and all in a group, but the rise is not linear, as some change their relative position. That is why this property is really important and remarkable. By using the PageRank algorithm if there is variation, there are changes in the Ranking within the same layer.

As in [21], this model of linguistic network arises from the variable nature of the text, which is always moving forward, so that directed and weighted links are used to represent relationships between linguistic units as in [14,42,43]. The co-occurrence relationship is established between two adjacent words or linguistic units within a sentence, where the direction of the link reflects the sequence of the words, and the weight on the link reflects the frequency of the appearance of that sequence of two linked words.

At this point it is important to highlight that there are different approaches to analyze a language from the perspective of complex networks [14,28] but they are different from the one we are proposing here. In our model we consider a network built from the corpus under study: network nodes are the words that appear in any of the texts that make up the corpus, establishing a (directed) link connecting two words if they appear consecutively (directed co-occurrence) somewhere in a text. In addition, we place the nodes in four different layers, in which the layers have a different meaning than the one used in [42]:

1. A layer in which we have included the specific words (mainly adjectives and nouns) of the specialty language, which we call the *lexical layer*. In this layer, we have included the terms that

are exclusive of the specialty mathematical language together with the terms subject to refer to different concepts when they don't show in mathematical context. We consider all this terms to be "lexical units" or "relevant specific words".
2. The *verb layer*, where we have included all the verbs regardless of their conjugation.
3. A layer in which we place the linking words, the *linking layer*. In this layer we have included words with gramatical relevance in terms of sentence construction and linking regarding both sentence correction and grammar correction. Prepositions, connectors, frequency adverbs and determiners respond to the this description.
4. And finally, a layer in which we place the remaining (uncharacterized) words, the *remaining words*, where we have included the words which do not suffer any alteration when they show in mathematical text because either they are words of the everyday language (non mathematical) or they are words of another field of specialty to which the complex network theory is being applied.

So, we have a directed network $G = (N, E)$ with four layers, in which N is the set of different words of the text ($N = V_1 \cup V_2 \cup V_3 \cup V_4$) and E is the set of directed (intra and interlayer) edges.

Finally, to complete this section, it should be noted that the development of tools for searching for lexical patterns in a speciality language, as well as the automatic classification of texts and the automatic extraction of significant texts from a corpus are some possible applications of the methodology underlying this model.

3. A new structure: Enriched line graph of a weighted network

When analyzing systems of different nature represented by networks it is important to point out the most relevant components, regardless of whether they are nodes or edges, or even small relevant parts of components of a subsystem of the system under analysis (for example when studying the cyber security of a computer network [13,45], or when working with language networks [21,42]). As it is known, the concept of line graph was introduced in [53] and since then this structure has been used to study certain properties of networks where the edges or connections between nodes may occasionally be more relevant than nodes themselves [1,5,33]. However, and quite surprisingly, when studying networks with a significant number of nodes and edges, line plots have only been considered in a reduced number of studies and applications [15,16,18,19,22,25,26,48]. In this section, given a network weighted directed network $G = (X, E, f)$, we will introduce the concept of Enriched line graph of threshold $\theta \in [0, +\infty)$, denoted by ELG^θ and its associated semi line graph SLG^θ that will allow us to describe in a specific way these components with a view to their applications.

Let $G = (X, E, f)$ be a *weighted directed network*, where $X = \{1, \ldots, n\}$ is the set of *vertices* or *nodes*, $E \subseteq X \times X$ is the set of *edges* and f is a function $f : E \longrightarrow (0, +\infty)$ which represents some kind of flow (electricity, data, water or similar) that circulates between two nodes through the edge that joins them, in such a way that for each edge $(i, j) \in E$, the coefficient $f(i, j)$ is called *weight* or *flow* of (i, j). We will also use the notation $i \to j$ for an edge (i, j) when convenient. Throughout this paper we will consider simple networks without loops, i.e., for every $i \in X$ we have that $(i, i) \notin E$, and also without multiple edges. We also consider the (in and out) neighbors of a node $i \in X$: $N^+(i) = \{j \in X | (j, i) \in E\}$, $N^-(i) = \{j \in X | (i, j) \in E\}$ and $N(i) = N^-(i) \cup N^+(i)$.

Typically, a network will have only a few nodes where flow enters or leaves the network; at all other nodes, the total incoming flow to a node is equal to the total flow out. So, for our model this means that, at most nodes $i \in X$, the function f satisfies *Kirchhoff's law*, i.e.,

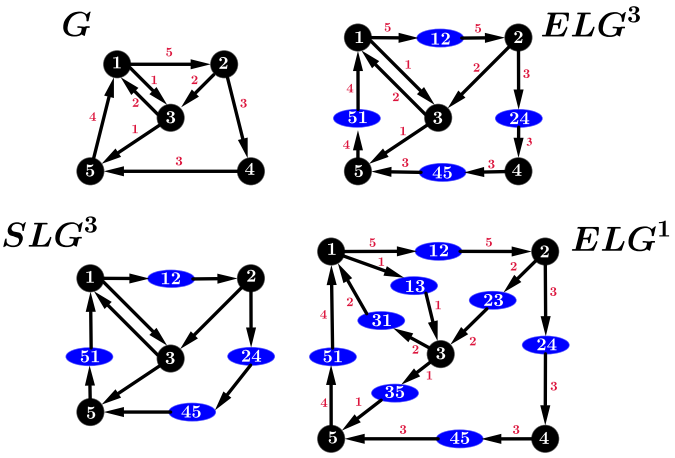

Fig. 3. An example on a specific network of its associated graphs ELG^3, SLG^3 and ELG^1. In this example, because the flow of all the edges is ≥ 1, $SLG^1 = LG$ (the line graph of G).

$$\sum_{k \in N^+(i)} f(k \to i) = \sum_{j \in N^-(i)} f(i \to j).$$

Given a directed and weighted network $G = (X, E, f)$, the *(weighted) adjacency matrix* of G is the matrix $A(G) = A = (a_{ij}) \in M_{n \times n}$ given by

$$a_{ij} = \begin{cases} f(i, j), & \text{if there exists an edge } (i, j) \in E, \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

Definition 3.2. Let $G = (X, E, f)$ be a weighted network (directed or undirected) and a positive real number $\theta \in (0, +\infty)$. We define the Enriched line graph of threshold θ associated to G as the weighted network $ELG^\theta = (X \cup E^\theta, E'_\theta, F)$ determined by the following conditions (see Fig. 3):

$$E^\theta = \{l \in E | f(l) \geq \theta\},$$

$$E'_\theta = E - E^\theta \cup \{(i, (i, j)) | (i, j) \in E \wedge f(i, j) \geq \theta\} \cup \{((i, j), j) | (i, j) \in E \wedge f(i, j) \geq \theta\}$$

and $F : E'_\theta \longrightarrow [0, +\infty)$ defined by

$$F(a, b) = f(i, j)$$

if $(a, b) = (i, j)$ or $(a, b) = (i, (i, j))$ or $(a, b) = ((i, j), i)$.

Definition 3.3. Let $G = (X, E, f)$ be a weighted network (directed or undirected), a positive real number $\theta \in (0, +\infty)$ and $ELG^\theta = (X \cup E^\theta, E'_\theta, F)$ the corresponding enriched line graph of G. We will say that a node $u \in X$ is a *dispensable node* of ELG^θ if $\forall v \in N(u)$ $\exists w \in X$ such that $v = (u, w) \vee v = (w, u)$.

In the sequel we will denote the set of dispensable nodes of ELG^θ by $Dis(ELG^\theta)$.

Definition 3.4. Let $ELG^\theta = (X \cup E^\theta, E'_\theta, F)$ be the enriched line graph of threshold $\theta \in (0, +\infty)$ of a given weighted network $G = (X, E, f)$. We define the Semi line graph of threshold θ, as the Unweighted network SLG^θ obtained by removing from ELG^θ the dispensable nodes $Dis(ELG^\theta)$ and the incident edges in them, and connecting the nodes affected by the elimination of one of their neighbors in the following way: If $i \in Dis(ELG^\theta)$ and $(i, j), (k, i) \in E^\theta$ then we add to SLG^θ the edge $((k, i), (i, j))$ (see Fig. 5).

Now, it is not difficult to prove the following theorem that shows that the semi line graph naturally extends the classic notion of line graph by considering the corresponding adjacency matrices:

Theorem 3.5. *Let $G = (X, E, f)$ be a directed strongly connected weighted network, where $f : E \longrightarrow [0, +\infty)$. Then*

$$\lim_{\theta \to 0^+} SLG^{\theta} = LG,$$

where LG is the classic line graph of G.

4. The multilayer enriched line graph

There are several ways to extend the line graph concept to multiplex and multilayer networks [19]. In our case, we are mainly interested in the structure of the enriched line graph resulting from considering only the nodes of a layer and their corresponding intra-layer edges. So, we consider a weighted (directed or undirected) multilayer network $\mathcal{M}$ [7], with m layers $\mathcal{M} = (\mathcal{G}, \mathcal{C})$ where $\mathcal{G} = \{G_\alpha;\ \ \alpha \in \{1, \cdots, m\}\}$ is a collection of (directed or undirected) weighted graphs $G_\alpha = (X_\alpha, E_\alpha)$, and

$$\mathcal{C} = \{E_{\alpha\beta} \subset X_\alpha \times X_\beta;\ \ \alpha, \beta \in \{1, \cdots, m\}, \alpha \neq \beta\}$$

is the set of interlinkages between nodes of different layers G_α and G_β with $\alpha \neq \beta$. The elements of $\mathcal{C}$ are called *crossed layers* and we also assume that they're all weighted. So, if $l \in E_{\alpha\beta}$ we will say that l is a *crossed edge* between the layers G_α and G_β. The set of nodes of $\mathcal{M}$ is

$$X = \bigcup_{\alpha=1}^{m} X_\alpha,$$

where $X_\alpha = \{x_1^\alpha, \ldots, x_{N_\alpha}^\alpha\}$ and $|X| = \sum_{\alpha=1}^{m} |X_\alpha| = N$. The set of intra-layers of $\mathcal{M}$ is $L = \{\ell_\alpha;\ \ \alpha \in \{1, \ldots, m\}\}$. Also, each layer of L is a directed and weighted graph $\ell_\alpha = (X_\alpha, E_\alpha)$ determined by the set of nodes $X_\alpha \subset X$ and the set of edges:

$$E_\alpha = \{e_{i,j}^\alpha; \alpha \in \{1, \ldots, m\}\},$$

where $e_{i,j}^\alpha$ represents the edge that connects nodes i and j and, as in the case of $\mathcal{C}$, we assume that they are all weighted.

Greek subscripts and superscripts are used to indicate the layer index so that N_α indicates the number of nodes of the layer ℓ_α (i.e., $N_\alpha = |X_\alpha|$). The adjacency matrix of layer ℓ_α is denoted by $A^{[\alpha]} = (a_{ij}^\alpha) \in \mathbb{R}^{N_\alpha \times N_\alpha}$, where

$$a_{ij}^\alpha = \begin{cases} f_\alpha(i,j), & \text{if there exists an edge } (i,j) \in E_\alpha, \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

and f_α is a function $f_\alpha : E_\alpha \longrightarrow [0, +\infty)$ such that for each edge $(i, j) \in E_\alpha$, the coefficient $f_\alpha(i, j)$ is called *weight* or *flow* of (i, j) in layer ℓ_α. Similarly, $\forall \alpha, \beta \in \{1, \ldots, m\}, \alpha \neq \beta$ we consider $f_{\alpha\beta} : E_{\alpha\beta} \longrightarrow [0, +\infty)$. In our model $f_\alpha(i, j)$ and $f_{\alpha\beta}(i, j)$ will represent the frequency of appearance in the corpus of the edge (i, j) (as intra-layer or as interlayer respectively).

Now, given a multilayer weighted network $\mathcal{M}$, as it is known, there are several monolayer networks associated with the multilayer network $\mathcal{M}$ that provide us with useful information about its structure. One of them is the *projection network* $\overline{\mathcal{M}} = (X, \overline{E})$ associated to $\mathcal{M}$, where X is the same set of nodes of $\mathcal{M}$ and $\overline{E}$ is the set of unordered pairs $e_{i,j} = i \to j$ such that $e_{i,j} \in E_\alpha$ for some $\alpha \in \{1, \ldots, m\}$ or $e_{i,j} \in E_{\alpha\beta}$ for some $\alpha, \beta \in \{1, \ldots, m\}, \alpha \neq \beta$.

It is clear that if $\overline{A} = (\overline{a}_{ij}) \in \mathbb{R}^{N \times N}$ is the adjacency matrix of $\overline{\mathcal{M}}$, then

$$\overline{a}_{ij} = \sum_{\alpha=1}^{m} f_\alpha(i,j) + \sum_{\alpha,\beta=1,\alpha\neq\beta}^{m} f_{\alpha\beta}(i,j).$$

.

For the rest of this section, as in the mono-layer case, we suppose that Kirchhoffs' law is satisfied, that is, there are only a few nodes where flow enters or leaves the network; at all other nodes, the total incoming flow the total flow incoming to a node is equal to the total flow out.

Definition 4.2. Let $\mathcal{M} = (\mathcal{G}, \mathcal{C})$ be a weighted (directed or undirected) multilayer network with m layers $\mathcal{M} = (\mathcal{G}, \mathcal{C})$ where $\mathcal{G} = \{G_\alpha;\ \ \alpha \in \{1, \cdots, m\}\}$ and a positive real number $\theta \in (0, +\infty)$. We define the enriched line graph of threshold θ associated to $\mathcal{M}$ as the enriched line graph $EL\overline{\mathcal{M}}^{\theta}$ associated to $\overline{\mathcal{M}}$. In a similar way, we define the Semi line graph associated to $\mathcal{M}$ of threshold θ, as the weighted network $SL\overline{\mathcal{M}}^{\theta}$.

Definition 4.3. Let $\mathcal{M} = (\mathcal{G}, \mathcal{C})$ be a weighted (directed or undirected) multilayer network with m layers $\mathcal{M} = (\mathcal{G}, \mathcal{C})$ where $\mathcal{G} = \{G_\alpha;\ \ \alpha \in \{1, \cdots, m\}\}$ and a positive real number $\theta \in (0, +\infty)$. We define the enriched line graph of threshold θ associated to the layer G_α as the weighted network $ELG_\alpha^\theta = (X \cup E_\alpha^\theta, E'_{\alpha-\theta}, F_\alpha)$ determined by the following conditions:

$$E_\alpha^\theta = \{l \in E_\alpha | f_\alpha(l) \geq \theta\}$$

$$\begin{aligned} E'_{\alpha-\theta} = {} & E_\alpha - E_\alpha^\theta \cup \{(i, (i, j)) | (i, j) \in E_\alpha \wedge f_\alpha(i, j) \geq \theta\} \\ & \cup \{((i, j), j) | (i, j) \in E_\alpha \wedge f_\alpha(i, j) \geq \theta\} \end{aligned}$$

and $F_\alpha : E'_{\alpha-\theta} \longrightarrow [0, +\infty)$ defined by

$$F_\alpha(a, b) = f_\alpha(i, j)$$

if $(a, b) = (i, j)$ or $(a, b) = (i, (i, j))$ or $(a, b) = ((i, j), j)$.

As it may seen in the previous definition, with a view to applications, we are interested in the enriched line graph of a specific layer, ignoring, in principle, the crossed edges of the multilayer network. In the specific case of the multilayer linguistic network under study, the layer that is a priori more interesting to study is the lexical layer. The definition of dispensable node of $ELG_\alpha^\theta = (X \cup E_\alpha^\theta, E'_{\alpha-\theta}, F_\alpha)$ is a direct translation from the monolayer case. Now, the definition of the semi line graph in this context is the following:

Definition 4.4. Let $ELG_\alpha^\theta = (X \cup E_\alpha^\theta, E'_{\alpha-\theta}, F_\alpha)$ be the enriched line graph of threshold $\theta \in (0, +\infty)$ associated to a given layer G_α of a weighted multilayer network $\mathcal{M} = (\mathcal{G}, \mathcal{C})$. We define the Semi line graph associated to G_α of threshold θ, as the weighted network SLG_α^θ obtained by removing from ELG_α^θ the dispensable nodes ELG_α^θ) and the incident edges in them, and connecting the nodes affected by the elimination of one of their neighbors in the following way: If $i \in Dis(ELG_\alpha^\theta)$ and $(i, j), (k, i) \in E_\alpha^\theta$ then we add to SLG_α^θ the edge $((k, i), (i, j))$.

Similarly to the previous case, and using the previously established notation, it is not difficult to prove the following result that relates the new concept with the classic line graphs of $\overline{\mathcal{M}}$ and G_α:

Theorem 4.5. *Let $\mathcal{M} = (\mathcal{G}, \mathcal{C})$ be a directed weighted multilayer network such that its projection network $\overline{\mathcal{M}} = (X, \overline{E})$ is a directed strongly connected weighted network. Then we have that*

$$\lim_{\theta \to 0^+} SL\overline{\mathcal{M}}^{\theta} = L\overline{\mathcal{M}}$$

where $L\overline{\mathcal{M}}$ is the line graph of $\overline{\mathcal{M}}$. Similarly, if G_α is a directed strongly connected weighted layer of $\mathcal{M}$, then

$$\lim_{\theta \to 0^+} SLG_\alpha^\theta = LG_\alpha$$

where LG_α is the line graph of G_α.

5. Computational results, rankings and applications to the search for collocations

In order to perform our study on the multilayer linguistic network considered, we are interested in a special type of flow that associates each node and edge of the network with its corresponding personalized PageRank [8,11,12,36,47].

As it is known, the personalized PageRank of a individual term (node) i is the i-component of the stationary state $\pi_0 \in \mathbb{R}^n$ ($\|\pi_0\| = 1$) of the random walker with transition matrix

$$P = \alpha P_A^T + (1-\alpha)\mathbf{v}\mathbf{e}^T,$$

where $\alpha \in (0,1)$, $\mathbf{e}^T = (1,\cdots,1)$, $\mathbf{v} \in \mathbb{R}^n$ ($\|\mathbf{v}\| = 1$) is the personalization vector and

$$P_A = (p_{ij}) = \left(\frac{a_{ij}}{\sum_k a_{ik}}\right).$$

In order to obtain the personalization vector of the individual terms, we have computed the frequency of appearance of each of the words that appear at the beginning of each sentence of the corpus, to keep it in mind when defining the corresponding PageRank personalization vector used [20,31,47]. Thus, the corresponding component of the personalization vector of a specific term is the result obtained by adding 1 to its relative frequency of appearance at the beginning of the sentences of the corpus texts, and dividing it by the number of nodes. In this way, if v is the personalization vector, its components are, $\forall i \in \{1,\ldots,n\}$,

$$v_i = \frac{1+f_i}{n+\sum_i f_i},$$

where f_i is the frequency described above. On the other hand, for the computation of PageRank used throughout this work we have used the algorithm described in [4].

As in [21], having in mind that the average number of words of a sentence within the corpus under study is 18.44, in this context, the damping factor corresponding to this configuration is 0.94, since

$$\begin{aligned} 18.44 = \mathbb{E}(\ell) &= \textstyle\sum_{k=0}^{\infty} k\cdot\mathbb{P}(\ell=k) = \sum_{k=1}^{\infty} k\cdot(1-q)\cdot q^k \\ &= (1-q)\cdot q\textstyle\sum_{k=1}^{\infty} k\cdot q^{k-1} = \frac{q}{1-q}. \end{aligned}$$

Hence, in this situation the damping factor to apply is $q = 0.94$. This value of q is the probability that a random walker will not vary its trajectory by moving to a node directly connected by an edge to the current node instead of jumping to another node in this network not connected to the previous one. In our situation, this jump can be understood as the end of the current sentence and the starting point of a new sentence.

Now, it is relevant to point out the known fact that there is a strong relationship between the personalized PageRank of the nodes and that of the edge that connects them [20]. In fact, the PageRank of a node i, $PR(i)$ depends essentially on the PageRank of the edges that coming out from i (Fig. 4), since the PageRank of the edges that come out from node i depend essentially on the PageRank of the edges that arrive at node i.

It is important to have in mind, as in [20], that the random jump by a random walker over the network can be interpreted as the disappearance of that random walker in a certain node and the appearance of a new random walker in another node and that, according to the PageRank vision within a biplex approach so that in one of the two layers of this vision we have a complete graph, as presented in [47], we can consider our managed network as a *strongly connected network*. In this way, to calculate the PageRank of both nodes and edges of the network we can think of thousands of random walkers moving over the network to calculate the flow corresponding to that node or edge (in fact, its PageRank).

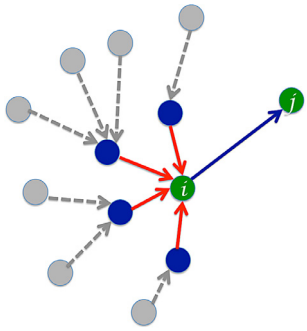

Fig. 4. $PR(i)$ depends essentially on the PageRank of edges that coming out from i.

Now, in order to carry out an efficient search for the combinations of words in the lexical layer that truly constitute collocations, we consider the enriched graph associated to the lexical layer (whose vertices are words from the lexical layer or combinations of two words from that lexical layer, i.e., intra-layer connections of lexical layer). We can take into account the PageRank variation of an "i" node when a combination of two words from the lexical layer in which he is the initial node is introduced, according to the expression:

$$PR^n(i) = PR^a(i) - PR(i \to j),$$

where $PR^n(i)$ is the new PageRank of node i, $PR^a(i)$ is the previous PageRank of i, and $PR(i \to j)$ is the PageRank of the edge $(i \to j)$. Now, by successively adding to the enriched line graph the nodes corresponding to the combinations of two words in the lexical layer, the search for collocations is optimized by considering the case in which $PR(i \to j) > PR^n(i)$.

Now we are going to see different ways to go adding new nodes corresponding to intra-edges of the lexical layer $G_1 = (X_1, E_1)$ until obtaining the enriched line graph ELG^1 by observing the variation that takes place in the Kendall's τ coefficient according to the criterion of addition of new nodes (in fact, intra-edges of G_1) considered. So, it is clear that the incorporation of these nodes contributes to the diminution of the PageRank of certain nodes of G_1 and to important variations in the ranking of these nodes ordered by the value of their PageRank.

In Fig. 5 it can be observed the Kendall's τ coefficient variation in the ranking of the first 500 nodes of the lexical layer by introducing the 1000 top edges (pairs of consecutive terms of the corpus) in the complete multilayer network according to different criteria: adding successively to the mixed line graph the best PageRank collocations (red color), adding successively the worst PageRank collocations (blue color) adding the collocations that minimize (green color) or maximize (purple color) the Kendall's τ coefficient, or adding collocations randomly (yellow color).

In a similar way, in Fig. 6 it can be observed the Kendall's τ coefficient variation in the ranking of the first 500 nodes of the lexical layer by introducing the 2726 collocations in the lexical layer according to the different criteria: adding successively to the mixed line graph the best PageRank collocations (red color), adding successively the worst PageRank collocations (blue color) adding the collocations that minimize (green color) or maximize (purple color) the Kendall's τ coefficient, or adding collocations randomly (yellow color).

The numerical experiments were run on a iMac18,3 with 4,2 GHz Intel Core i7 and RAM 16 GB, under the macOS High Sierra operating system. All the experimental results were obtained by using a Python 3.7 implementation with machine precision $\varepsilon \approx 2.22 \times 10^{-16}$.

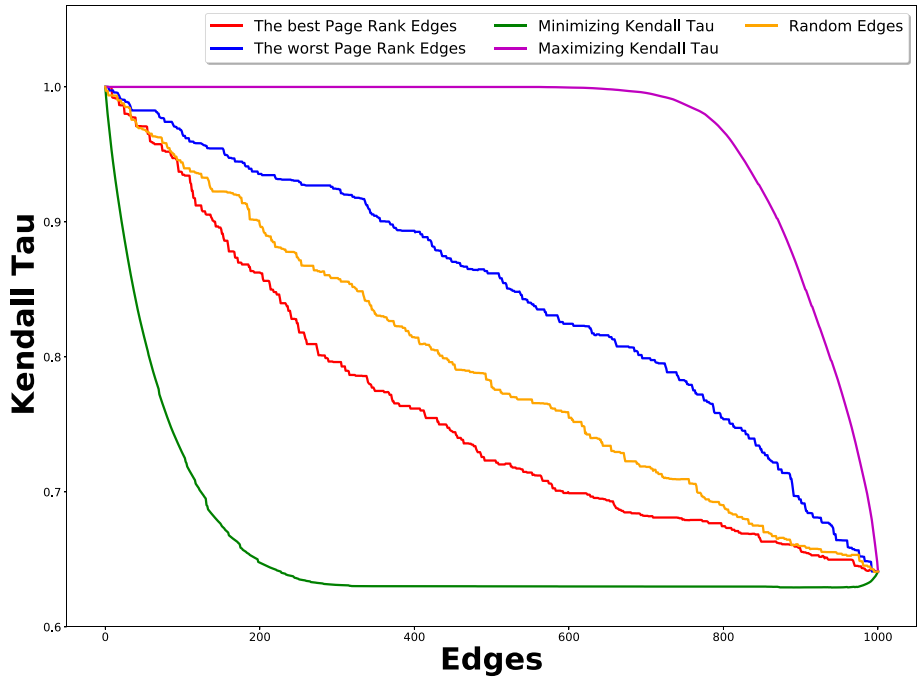

Fig. 5. Kendall's τ coefficient variation in the ranking of the 500 top words by introducing the 1000 top edges (pairs of words) according to different criteria.

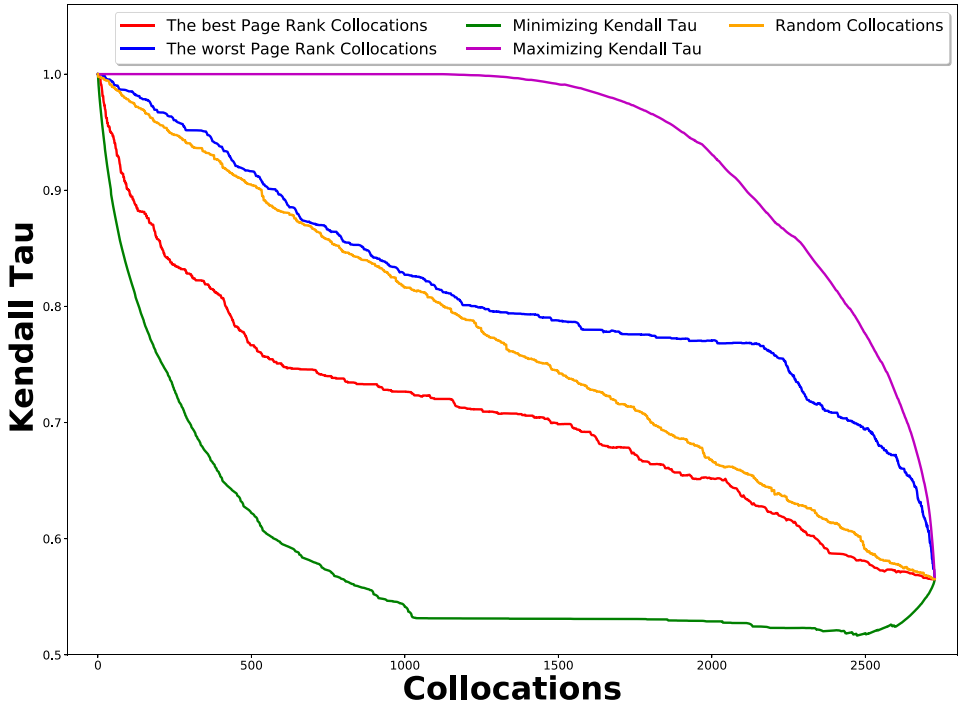

Fig. 6. Kendall's τ coefficient variation in the ranking of the first 500 nodes of the lexical layer by introducing the 2726 collocations in the lexical layer according to different criteria.

6. Seeking characteristics for the specialty language generated by the scientific community of complex networks

Once the new structures introduced (Enriched line graph and semi line graph) have been used to incorporate the collocations as new nodes of the multilayer network within the lexical layer, it is possible to obtain what we could call a "zipper ranking" that presents in the same list both the words and the collocations detected with the highest PageRank. This list of words and collocations is a unique feature of the speciality language analyzed. The optimal threshold θ for obtaining the Enriched line graph and the corresponding Semi line graph, with the aim of incorporating the collocations in our network, depends on the number of vertices and the edges weight distribution (numerical value of the PageRank). In the studied case, as it is natural, we have used some approximate values for θ close to a point where a phase transition occurs. Specifically, in Fig. 7 the following values have been explicitly represented: $\theta = 3.7208540316339136e-$

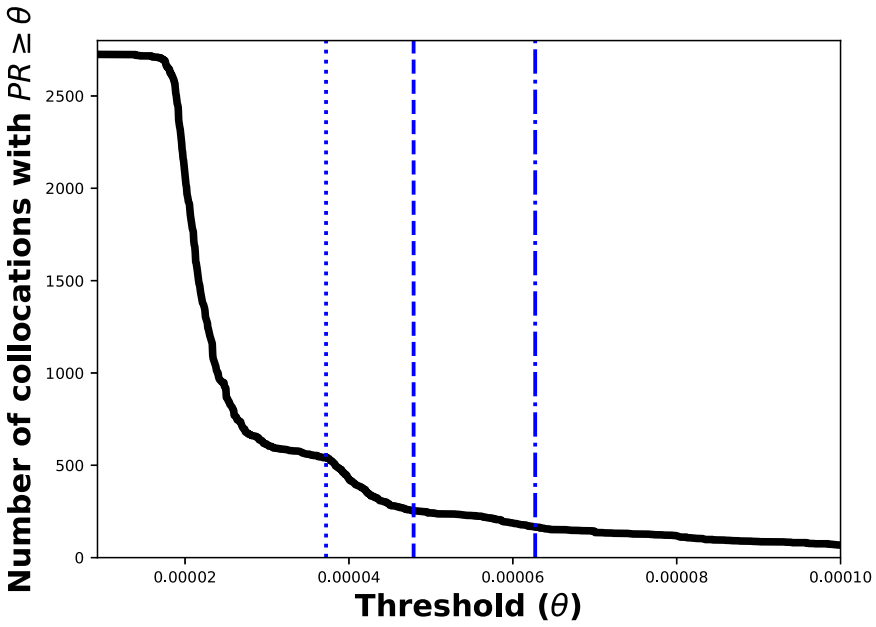

Fig. 7. Number of collocations obtained in the lexical layer for values of PageRank above the threshold θ.

Table 1
The top 10 words of the whole network, the lexical layer, the consecutive words and the "collocations".

	Whole network	Lexical layer	Consecutive words	Collocations
1st	the	network	*of → the*	*complex → networks*
2nd	of	networks	*the → network*	*complex → network*
3nd	to	nodes	*in → this*	*multiplex → network*
4th	and	system	*of → a*	*degree → distribution*
5th	in	information	*is → a*	*network → topology*
6th	a	number	*in → a*	*complex → systems*
7th	is	model	*is → the*	*topological → properties*
8th	that	structure	*the → same*	*network → structure*
9th	are	systems	*in → order*	*network → analysis*
10th	we	properties	*the → number*	*centrality → measures*

05 (dotted line), $\theta = 4.789277532128921e-05$ (dashed line) and $\theta = 6.27319906059421e-05$ (dotted and striped line). It is important to note that for the three values of *theta* represented, the first collocation ("complex networks") appears in 61*st* place for $\theta = 3.7208540316339136e-05$, in 64^{th} place for $\theta = 4.789277532128921e-05$, and in 66^{th} place for $\theta = 6.27319906059421e-05$. The following collocation appears for these three values of θ, respectively, in the places 83, 89 and 95. It is also important to point out that for a low threshold such as $\theta = 1.0497952803963938e-05$ it is necessary to reach the position 220^{th} of the full lexical layer zipper ranking ranking to have the first 10 collocations within the "zipper ranking".

In Table 1 the ranking of the ten words with the highest PageRank of the entire multilayer network is presented, as well as the ranking of the ten words with the highest PageRank within the lexical layer, joint to the collocations of lexical layer ordered by the same criteria and the consecutive word pairs of the entire multilayer network.

The order in which the collocations in the zipper ranking appear gives us an idea of the importance of these collocations in the speciality language considered. So Table 1 together the zipper ranking which indicates the place where the first collocations appear, constitute specific features of the specialty language used by the scientific community of complex networks and constitute a very valuable contribution to the correct translation of texts belonging to this field of science.

7. Conclusions

We introduced and studied Enriched Line Graph (ELG) and Semi Line Graph (SLG) as two new and useful structures that can be interpreted as a certain types of "interpolation" between the original graph and its associated line graph, showing some characteristics and properties of these structures and one specific application of these concepts to obtain some technical characteristics of the specialty language produced by the scientific community of complex networks. We have extended these concepts to the context of multilayer networks, and we have studied and analyzed the properties and relationships between them, with a view to classify the collocations in the lexical layer of the linguistic model presented. Some relationships of these new structures with the classic line graph structure are also established. Furthermore, our numerical experiments showed some specific properties about the differences between the rankings provided when we applied them to different layers of our network. Undoubtedly, the instruments derived from the linguistic analysis arising from this model will not only contribute to facilitating the correct translation of specialized languages; furthermore, it will provide enhanced tools to typifying and localizing the characteristics of specialized texts, such as their classification by area, author or style, among others. So, among the conclusions of this work, it is important to highlight that, although among the applications of this study we have focused fundamentally on the intra-layer links of the lexical layer, the study of other

types of links will allow us to go deeper into the linguistic study of the construction/formation of different speciality languages and to seek the answer to other questions of a linguistic nature such as"do speciality verbs exist?" (to study the relationships within the verbal layer) or "do certain word combinations only go with certain verbs?" (to study the inter-layer relations between the lexical layer and the verbal layer). Finally, it is important to mention that the construction of tools to find lexical patterns of a specialty language, the automatic extraction of meaningful texts from a corpus, and the automatic classification of texts are among the possible applications of this methodology, extensible to other specialty languages. Also, labelling and automatic identification/verification of texts can be enhanced with a more precise identification of lexical patterns such as collocations. The application of this methodology and new algorithms and tools to more specific and larger linguistic corpora with a view to a more precise characterization of specialty languages is part of our future work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been partially supported by projects PGC2018-101625-B-I00 (Spanish Ministry, AEI/FEDER, UE) and M1993 (URJC Grant).

References

[1] Aigner M. On the line graph of a directed graph. Math Z 1967;102(1):56–61.
[2] Albert R, Barabasi AL. Statistical mechanics of complex networks. Rev Mod Phys 2002;74:47–97.
[3] Albert R, Jeong H, Barabasi A. Diameter of the world-wide web. Nature 1999;401:130–1.
[4] Aleja D, Criado R, García del Amo A, Pérez A, Romance M. Non-backtracking pagerank: from the classic model to Hashimoto matrices. Chaos, Solitons Fractals 2019;126:283–2918.
[5] Bagga J. Old and new generalizations of line graphs. IJMMS 2004;29:1509–21.
[6] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU. Complex networks: structure and dynamics. PhysRep 2006;424:75–308.
[7] Boccaletti S, Bianconi G, Criado R, Del Genio CI, Gómez-Gardeñes J, Romance M, Sendiña Nadal I, Wang Z, Zanin M. The structure and dynamics of multilayer networks. Phys Rep 2014;544(1):1–122.
[8] Boldi P, Santini M, Vigna S. Pagerank: functional dependencies. ACM Trans Inf Syst 2009;27:419–23.
[9] Borge-Holthoefer J, Arenas A. Semantic networks: structure and dynamics. Entropy 2010;12:1264–302.
[10] Bowker L, Pearson J. Working with specialized language: a practical guide to using corpora. Routledge; 2002.
[11] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Comput Netw 1998;30:107.
[12] Brin S, L P, Motwani R, Winograd T. The pagerank citation ranking: bringing order to the web. Standford InfoLab; 1999. Technical report.
[13] Chapela V, Criado R, Moral S, Romance M. Intentional risk management through complex networks analysis. Heidelberg New York Dordrecht London: Springer International Publishing; 2015.
[14] Cong J, Liu H. Approaching human language with complex networks. Phys Life Rev 2014;11(4).
[15] Criado R, Flores J, García del Amo A, Romance M. Analytical relationships between metric and centrality measures of a network and its dual. JCAM 2011;235(7):1775–80.
[16] Criado R, Flores J, García del Amo A, Romance M. Structural properties of the line-graphs associated to directed networks. Netw Heterogen Media 2012;7(3):373–84.
[17] Criado R, Flores J, García del Amo A, Gómez-Gardeñes J, Romance M. A mathematical model for networks with structures in the mesoscale. Int J Comput Math 2012;89(3):291–309.
[18] Criado R, Flores J, García del Amo A, Romance M. Centralities of a network and its line graph: an analytical comparison by means of their irregularity. IntJComputMath 2014;91(2):304–14.
[19] Criado R, Flores J, García del Amo A, Romance M, Barrena E, Mesa JA. Line graphs for a multiplex network. Chaos 2016;26(6):065309.
[20] Criado R, Moral S, Pérez A, Romance M. On the edges's pagerank and line-graphs. Chaos 2018;28(7):075503.
[21] Criado-Alonso A, Battaner-Moro E, Aleja D, Romance M, Criado R. Using complex networks to identify patterns in specialty mathematical language: a new approach. Soc Netw Anal Min 2020;10(1):1–10.
[22] Crucitti P, Latora V, Porta S. Centrality in networks of urban streets. Chaos 2006;16(015113).
[23] Da Fontoura Costa L, Oliveira ON, Travieso G, Rodrigues FA, Villas Boas PR, Antiqueira L, et al. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. Adv Phys 2011;60(3):329–412.
[24] Dogorovtsev SN, Mendes JFF. Language as an evolving word web. Proc R Soc Lond B 2001;268:2603–6.
[25] Evans TS, Lambiotte R. Line graphs, link partitions, and overlapping communities. Phys Rev E 2009;80:016105.
[26] Evans TS, Lambiotte R. Line graphs of weighted networks for overlapping communities. Eur Phys J B 2010;77:265–72.
[27] Estrada E. Networks science. New York: Springer; 2010.
[28] Ferrer i Cancho R, Solé RV. The small world of human language. Proc Royal Soc London B 2001;286:2261–6.
[29] Ferrer i Cancho R, Riordan O, Bollobás B. The consequences of Zipf's law for syntax and symbolic reference. ProcBiol Sci/R Soc 2005;272(1562):561–5.
[30] Firth JR. Modes of meaning. In: Essays and studies of the english association, N.S., vol. 4; 1951. p. 118–49.
[31] García E, Pedroche F, Romance M. On the localization of the personalized pagerank of complex networks. Linear Algebra Appl 2013;439:640–52.
[32] Halliday M.A.K., Matthiessen C.M.I.M.. Introduction to functional grammar (third edition). Routledge, Taylor & Francis Group. London and New York, 2004,
[33] Hemminger RL, L W Beineke LW. Line graphs and line digraphs. Selected topics in graph theory. Lowell WB, Wilson RJ, editors. New York: Academic Press; 1978.
[34] Huddleston RD. The cambridge grammar of the english language. Cambridge, UK; New York: Cambridge University Press; 2002.
[35] Kivela M, et al. Multilayer networks. J Complex Netw 2014;2(3):203–71.
[36] Langville AN, Meyer CD. Google's pagerank and beyond: the science of search engine ranks. Princeton Univ Press; 2006.
[37] Latora V, Nicosia V, Russo G. Complex networks: principles, methods and applications. Cambridge University Press; 2017.
[38] Liu H, Hu F. What role does syntax play in a language network? EPL 2008;83:18002.
[39] Liu H, Cong J. Empirical characterization of modern chinese as a multi-level system from the complex network approach. J ChinLinguist 2014;42:1–38.
[40] Liu H, Xu C, Liang J. Dependency distance: a new perspective on syntactic patterns in natural languages. Phys Life Rev 2017;21:171–93.
[41] Mc Enery T, Hardie A. Corpus linguistics: method, theory and practice. Cambridge textbooks in linguistics. Cambridge: Cambridge University Press; 2011.
[42] Martincic S, Margan D, Mestrovic A. Multilayer network of language: a unified framework for structural analysis of linguistic subsystems. Phys Rev E 2016;74:026102.
[43] Masucci A, Rodgers G. Network properties of written human language. Physica A 2006;457:117–28.
[44] Mehler A, Lücking A, Banisch S, Blanchard P, Frank-Job B. Towards a theoretical framework for analyzing complex linguistics networks. Springer-Verlag; 2016.
[45] Moral S, Chapela V, Criado R, Pérez A, Romance M. Efficient algorithms for estimating loss of information in a complex network: applications to intentional risk analysis. Netw Heterogen Media 2015;10(1):195–208.
[46] Newman M. Networks: an introduction. Oxford University Press; 2010.
[47] Pedroche F, Romance M, Criado R. A biplex approach to pagerank centrality: from classic to multiplex networks. Chaos 2016;26(6):065301.
[48] Porta S, Crucitti P, Latora V. The network analysis of urban streets: a primal approach. Environ Plann B 2006;33(5):705–25.
[49] Sinclair J. Corpus, concordance, collocation. Describing english language. Oxford University Press; 1991.
[50] Solé R. Syntax for free? Nature 2005;434:289.
[51] Solé RV, Corominas-Murtra B, Valverde S, Steels L. Language networks: their structure, function, and evolution. Complexity 2010;15(6):20–6.
[52] Wasserman S, Faust K. Social network analysis. Cambridge: Cambridge University Press; 1994.
[53] Whitney SH. Congruent graphs and the connectivity of graphs. Am J Math 1932;54(1):150–68.
[54] Zipf GL. Human behavior and the principle of least effort. Hafner; 1965.