Scientometrics (2012) 91:827–842
DOI 10.1007/s11192-012-0630-z

Using complex networks concepts to assess approaches for citations in scientific papers

D. R. Amancio · M. G. V. Nunes · O. N. Oliveira Jr. · L. da F. Costa

Received: 29 August 2011 / Published online: 28 January 2012
© Akadémiai Kiadó, Budapest, Hungary 2012

Abstract The number of citations received by authors in scientific journals has become a major parameter to assess individual researchers and the journals themselves through the impact factor. A fair assessment therefore requires that the criteria for selecting references in a given manuscript should be unbiased with regard to the authors or journals cited. In this paper, we assess approaches for citations considering two recommendations for authors to follow while preparing a manuscript: (i) consider similarity of contents with the topics investigated, lest related work should be reproduced or ignored; (ii) perform a systematic search over the network of citations including seminal or very related papers. We use formalisms of complex networks for two datasets of papers from the arXiv and the Web of Science repositories to show that neither of these two criteria is fulfilled in practice. By representing the texts as complex networks we estimated a similarity index between pieces of texts and found that the list of references did not contain the most similar papers in the dataset. This was quantified by calculating a consistency index, whose maximum value is one if the references in a given paper are the most similar in the dataset. For the areas of "complex networks" and "graphenes", the consistency index was only 0.11–0.23 and 0.10–0.25, respectively. To simulate a systematic search in the citation network, we employed a traditional random walk search (i.e. diffusion) and a random walk whose probabilities of transition are proportional to the number of the ingoing edges of the neighbours. The frequency of visits to the nodes (papers) in the network had a very small correlation with either the actual list of references in the papers or with the number of downloads from the arXiv repository. Therefore, apparently the authors and users of the repository did not follow the criterion related to a systematic search over the network of citations. Based on these results, we propose an approach that we believe is fairer for

D. R. Amancio (✉) · O. N. Oliveira Jr. · L. da F. Costa
Institute of Physics of São Carlos, University of São Paulo, P.O. Box 369,
São Carlos, São Paulo 13560-970, Brazil
e-mail: diegoraphael@gmail.com

M. G. V. Nunes
Institute of Mathematics and Computer Science, University of São Paulo,
P.O. Box 369, São Carlos, São Paulo 13560-970, Brazil

evaluating and complementing citations of a given author, effectively leading to a virtual scientometry.

Keywords Complex networks · Virtual scientometry · Similarity network

Introduction

The advance of knowledge is founded and critically dependent on the broad dissemination of novel approaches and results, which allows other scientists and practitioners to analyze reported results to validate and complement their investigations. The objective of any scientific publication, in addition to certificating/authenticaticating knowledge (De Mey 1982), is to be *read, tried*, and *cited* by as many people as possible (McClellan 2003). Indeed, articles have been evaluated in terms of the citations they motivate, while journals are typically rated according to the impact factors reflecting the number of citations to their articles. Citations have been a major factor since the 1920s (Gross and Gross 1927), and subjects such as this are now analyzed in scientometrics, which studies the relationship between areas of knowledge and the evolution of science (Börner et al. 2004). Of course, the success of a paper in being read and cited varies enormously owing to several factors, including the renown of the journal and the eminence of the authors and their institutions. Strictly speaking, such a success should depend not only on the quality, originality, completeness and clarity of a specific paper, but also on the degree of relationship and overlap with the investigation being reported. For all the papers strongly similar or related to a current investigation should be read, and potentially cited. However, with the limited time available to any researcher for seeking and reading, related works have to be filtered by using some limiting criteria. Though unavoidable, this implies that important publications are overlooked (Wright and Armstrong 2008; Lilien 2008), which may undermine the efficiency of the whole system, in the sense that painstaking, costly efforts are repeated or ignored.

We take the view that increasing attention should be given to the procedures of selecting publications for guiding the research and preparing a list of references (May 1967) as many patterns of citations go far beyond the recommended conventions of scholarly publishing (Bornmann and Daniel 2008). In this paper we analyze two criteria for assessing the quality of selected citations in scientific manuscripts. The first is that similar, strongly related works should be selected to be read and potentially cited, and the second is that the authors should do a systematic search over related publications and their citations. We check whether these criteria are fulfilled by using complex networks (Barabási 2009; Albert and Barabási 2002; Newman 2003) and natural language processing formalisms (for the use of complex networks in natural language processing, see Ferrer et al. 2001, 2004; Sigman and Cecchi 2002; Costa 2004). Two datasets containing 900 articles each for the areas of complex networks and graphenes were used to obtain two networks for each area: (i) the traditional citation networks, where each article is a node and citations become directed edges between them; and (ii) a network obtained by the overlap between the contents of pairs of articles. These networks are henceforth referred to as *citation* and *overlap*, being directed and undirected, respectively. To calculate the similarity between the articles, we used the text files submitted to the arXiv[1] repository. To recover the citations of the articles, we used the arXiv repository and the Web of Science[2]. As for the overlap network,

[1] http://arxiv.org.

[2] http://apps.isiknowledge.com/.

each article was modeled as a complex network in order to extract the similarity relationship. Moreover, we also computed traditional similarity index to show that the analysis performed in this paper is weakly dependent on the similarity measure. The model used to represent texts as complex networks (see "Methodology"), which basically connects adjacent words after a pre-processing step, was chosen because of its success in other studies in Natural Language Processing, such as automatic text assessment (Antiqueira et al. 2007), automatic summarization strategies (Antiqueira et al. 2009) and automatic machine translation assessment (Amancio et al. 2008, 2011). After defining these two networks one may quantify the number of (a) articles which are related and cited; (b) articles that are related but not cited; and (c) articles that are loosely related but are cited nonetheless. We shall show that the analysis of these numbers indicates that the similarity criterion for selecting references is not obeyed. We also perform a random walk through the citation networks to simulate a systematic search by an author, whose results are used to infer that the second criterion is not obeyed either. In addition to discussing the possible causes and implications of these results, we suggest a virtual citation approach, where citations are artificially established according the content similarity between papers and references. This was implemented to complement the relationships among articles, which gives rise to a virtual scientometry.

Methodology

In our experiments, the relationships (similarity and citation) between two articles are modeled in complex networks. A network is defined as a data structure comprising a set of nodes linked by edges. The set of edges and nodes can be represented as a matrix W_{ij}, where the presence of an edge between two nodes i and j with weight Ω leads to $W_{ij} = \Omega$ and the absence of an edge implies $W_{ij} = 0$. If there is no order distinction to link two nodes ($i \rightarrow j$ is the same as $j \rightarrow i$), then $W_{ij} = W_{ji}$ is always true. If two nodes are connected by an edge, they are said to be adjacent. If two edges are associated with the same node, they are called adjacent edges. A sequence of adjacent edges defines a walk over the network. The length of a walk is defined as the number of edges in the walk. The networks were built using a corpus comprising 900 articles about complex networks (or scale free networks) and 900 articles on graphenes from the arXiv publications base. The first 900 articles returned by the query in the arXiv database were selected to compose our database. The keywords employed in the query were "complex networks" or "scale-free networks" and "graphenes" in the title or in the abstract. Only the most current version of the manuscripts was considered. It should be mentioned that any search on the arXiv site returns a set of articles that are not necessarily in chronological order of publication.

The citation network W_{ij}^{cit}, which can be considered as a modified acyclic (Shevchuk and Snarskii 2010) social network (the relationship between people is replaced by the relationship between articles) (Patrick 1985) and is known to follow a power law distribution (Redner 1998), stores all the information about citations among articles, where each article is a node and each edge represents a citation. These networks have drawn the interest of many researchers. For instance, unbiased indexes were used to quantify the relevance of scientific works (Costa 2006) and citation networks were employed to reveal patterns of connectivity among authors in co-authorship networks (Cotta and Merelo 2005). In its simplest form, a citation network can be built as follows. If article i cites article j then there will be a directed edge network represented as $i \rightarrow j$. In case $i = j$, nothing is done, since we disregard citations for the very paper being produced. To construct the citation network,

in addition to the arXiv identifier, we collected in the reference list the title of the articles. This was done because articles already published are not cited using the arXiv identifier. Thus, using the relationship between title and arXiv identifier, we were able to recover most of the citations with no identifiers, but that in fact belong to the arXiv database. Moreover, we used the Web of Science database to get reliable citation information. We do not take into account the changes made in a document after it was posted on the arXiv, and appearing in a publication in a journal, for we only used the Web of Science to retrieve additional references. To illustrate a citation network, Fig. 1 shows the network obtained after analyzing the list of references of two papers, whose arXiv identifiers are 0902.3068 and 0805.2298.

For building the overlap network, each file representing a paper in the database was pre-processed to remove the tags from LaTeX markup language, so that specific terms from LaTeX do not affect the calculation of the similarity index (see below). The overlap network W_{ij}^{ov} is undirected and comprises vertices representing the articles (as in a network of citations) and edges that connect two articles with some degree of similarity. For each edge a weight proportional to the similarity between the two vertices is defined. To visualize the overlap networks we eliminated all weights below a threshold (see example in Fig. 3). For obtaining the similarity between two papers (nodes), a two-stage procedure was adopted: (i) modeling each text as a complex network (Antiqueira et al. 2005); and (ii) comparing the corresponding networks. In modeling the text as a network, the stopwords were removed. Moreover, the remaining words were lemmatized to combine concepts with the same canonical form, but with different inflections. Additionally, the text was labeled using the MXPost part-of-speech Tagger (Aires et al. 2000) based on the Ratnaparki's model (Ratnaparki 1997), which helps to resolve problems of ambiguity. This is useful because words with the same canonical form and same meaning are grouped into a single node, while words with the same canonical form but distinct meanings generate distinct nodes. This pre-processing is done by accessing a computational lexicon (Nunes et al. 1996), where each word has an associated rule for the generation of the canonical form. The structure representing the network derived from a text is a weighted adjacency matrix.

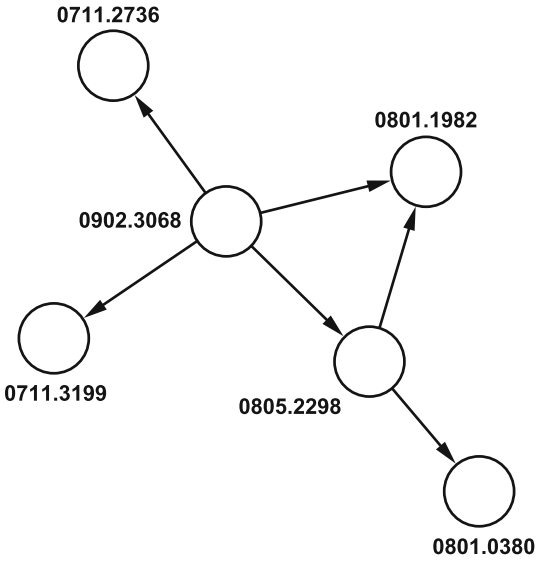


Fig. 1 Visualization of a citation subnetwork for the complex network subject. While article 0902.3068 cites four articles within the arXiv database, the article 0805.2298 cites only two articles within the same database

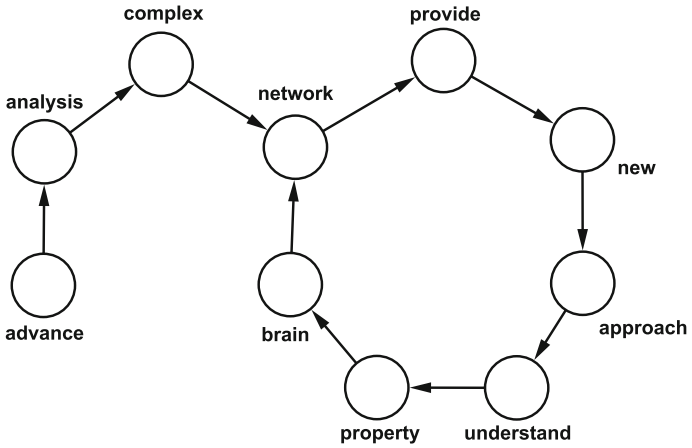


Fig. 2 Network obtained from the sentence "Advances in the analysis of complex networks provide useful new approaches to understanding many properties of brain networks"

After the pre-processing, the N words represented the network nodes and the resulting sequence of words was used to create the edges: for each pair of consecutive words there was a corresponding edge in the network. The edges also had weights, which indicated the number of times that the associations of words appeared in the text. This network was stored as a directed adjacency matrix, named W. The latter was initialized with zero elements, and as each word pair was read from the text, W_{ij} was incremented. In other words, the corresponding edge had its weight increased. Figure 2 shows the network from an adapted sentence extracted from the abstract of paper 0704.0392: "Advances in the analysis of complex networks provide useful new approaches to understanding many properties of brain networks".

Topology-based similarity index (TSI) and consistency index (CI)

It is well accepted that two pieces of text may be considered as similar if they share a large number of common concepts (represented by words), also taking into account the co-occurrence of concepts. To quantify the extent of similarity we propose an index calculated by comparing the networks of the texts considered, in which the metrics selected are believed to capture the network topology. Therefore, the index involves not only the common concepts (nodes) of the two networks, but also the vicinity of such nodes. Formally, the similarity index between two texts was obtained as follows. Let s be the number of occurrences of a word p in the first pre-processed text and t the number of occurrences of the same word in the second pre-processed text. The factor α_p defined as

$$\alpha_p = \frac{s + t - |s - t|}{s + t} \tag{1}$$

is a first measure of similarity. The reasoning behind this equation is quite simple. It quantifies the percentage of word frequency shared by two texts, since the numerator is twice the minimum among s and t and the denominator is twice the maximum shared frequency, which occurs when $s = t$.

We then use information from the adjacency word network. Let v_s be the node of the adjacency network representing the word p in the first pre-processed text and v_t the node of

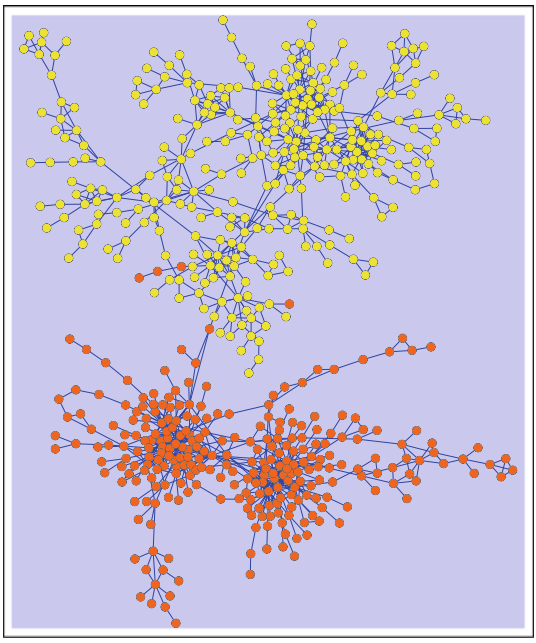


Fig. 3 Visualization of an overlapping network (largest component) for the articles belonging to two subjects: complex network (*dark tonality*) and graphene (*clear tonality*). The presence of only a few connections between vertices of different colors (different subjects) shows that the similarity index connects similar articles

the adjacency network for the same word in the second pre-processed text. If node v_s has κ neighbors in common with the neighbors of v_t, with W_s representing the first network and W_t the second, the β_p factor is defined as in Eq. 2. Similarly to Eq. 1, the numerator quantifies the double of the weight that is shared (i.e., the double of the minimum between $W_s(v_s, v)$ and $W_t(v_t, v)$, where v represents an arbitrary neighbor of v_s or v_t) and the denominator quantifies the double of the maximum number of shared weight. Therefore, like α_p, β_p is limited to the interval between 0 and 1.

$$\beta_p = \frac{\sum_{v=1}^{\kappa} \left(W_s(v_s, v) + W_t(v_t, v) - |W_s(v_s, v) - W_t(v_t, v)|\right)}{\sum_{v=1}^{\kappa} W_s(v_s, v) + \sum_{v=1}^{\kappa} W_t(v_t, v)} \tag{2}$$

With these definitions, the similarity index for the word p, referred to as I_p, is given in Eq. 3:

$$I_p = \alpha_p \beta_p \tag{3}$$

The global similarity index TSI between these two texts is then defined in Eq. 4, where N is the number of distinct words in the set formed by adding the sets of words in the two texts.

$$\text{TSI} = \frac{1}{N} \sum_{p=1}^{N} I_p \tag{4}$$

Assessing similarity between texts is highly subjective. Nevertheless, we believe that the index above is at least reasonable, as indicated in Fig. 3, where a visualization using the software Cytoscape[3] is provided for two distinct subjects: graphenes and complex networks. When linked as a single network with the edges between vertices being dependent

[3] http://www.cytoscape.org.

on the similarity index, the vertices of the same color remain practically separated from the other color. Since the color represents a given subject, and little overlap is seen, an effective separation is inferred via the similarity index.

In addition to the similarity index proposed in this paper, we employed 3 other traditional measures of similarity between pairs of texts: Jaccard, Cosine and Dice (Tan et al. 2005). The procedure to compute similarity is different for these three measurements, but they are all related because they are based on the frequency of terms. In other words, if two texts share many terms, their similarity will be high.

Because in this paper we consider that one of the criteria to consult and potentially cite a given paper should be based on the similarity of contents with the manuscript being written, we quantified whether this expectation is fulfilled by using a consistency index. By consistent we mean that while selecting papers to cite, the authors considered the similarity of content. The consistency index CI was conceived to vary from 0 to 1, where 1 corresponds to the case where all the references cited in the papers of the database are contained in the selected database (see the definition below). The value 0 is for the case where the citations in the papers do not belong to the selected database. We therefore refer to the total database as that containing the η papers from arXiv used in the citation network (we used $\eta = 900$), while the selected database comprises the papers cited which belong to the total database. The procedure to obtain CI is as follows. For each paper i in the database, one obtains the similarity threshold for which the number of similar papers in the database is equal to the number of references in paper i. The papers from the arXiv database satisfying this requirement for all the papers from $i = 1$ to $i = \eta$ comprise the so-called selected database. To illustrate: suppose paper i has 20 references. This paper will contribute with 20 papers—the most similar to paper i—to the selected database. Let x_i be the number of references of paper i which belong to the selected database. Let y_i be the number of references of paper i which belong to the complete arXiv database. The ratio x_i/y_i therefore gives the fraction of papers cited that are actually similar to the original one. For instance, suppose that 8 out of the 20 references in paper i belong to the arXiv database (complete database with η papers). Then, $y_i = 8$. If 4 of these 8 references appear in the selected database (the most similar given the threshold mentioned above), then $x_i = 4$, and the ratio is 0.5. The consistency index will be the sum of x_i/y_i, from i = 1 to η, divided by η, as shown in Eq. 5. High values will mean that the references are indeed similar to the original paper. We should stress that in selecting the most similar papers from the databases, we only considered those published before the manuscript under consideration, since concomitant or later papers could not be cited in practice.

$$\mathrm{CI} = \frac{1}{\eta}\sum_{i=1}^{\eta}\frac{x_i}{y_i} \tag{5}$$

Now, it is possible that CI = 1 and the similarity is not high. Let us exemplify with an extreme case. Suppose a paper with $x_i = 3$ and $y_i = 3$, but the 3 references are the least similar in the selected database (just above the threshold). In order to account for that we created another parameter CI′, which considers the order in which the papers appear (i.e. 1 for the most similar and 0 for the least similar—the last included in the selected database).

CI′ is defined in Eq. 6, which amounts to introducing a weight w_i into CI, depending on the rank r_k of each of the x_i articles in the list of similarity comprising y_i papers. This weight is computed so that higher ranks provide higher weights and vice-versa. To know the variation of the sum of r_k can get, one calculates first the maximum value $\Gamma_{\max}$ (it occurs when all the x_i papers are in the tail of the list) and the minimum value $\Gamma_{\min}$ (it occurs when all the x_i papers

are in the head of the list) as shown in Eqs. 7 and 8, respectively. Then, the range of the sum of r_k is computed in Eq. 9. Also, the difference between the sum of r_k and $\Gamma_{\min}$ is computed to quantify how far is this sum from the worst case (i.e., $\Delta\Gamma$). Finally, given these values, we are able to define the weight w_i according to Eq. 10.

$$\mathrm{CI}' = \frac{1}{\eta}\sum_{i=1}^{\eta}\frac{x_i\omega_i}{y_i} \tag{6}$$

$$\Gamma_{\min} = \sum_{k=1}^{x_i} k = \frac{x_i(x_i+1)}{2} \tag{7}$$

$$\begin{aligned}\Gamma_{\max} &= \sum_{k=0}^{x_i-1}(y_i-k) = x_i y_i - \sum_{k=0}^{x_i-1} k \\ &= \frac{2x_i y_i - x_i(x_i-1)}{2} = \frac{x_i}{2}(2y_i - x_i + 1)\end{aligned} \tag{8}$$

$$\begin{aligned}\Delta\Gamma &= \Gamma_{\max} - \Gamma_{\min} \\ &= \frac{x_i}{2}(2y_i - x_i + 1) - \frac{x_i}{2}(x_i+1) \\ &= x_i(y_i - x_i)\end{aligned} \tag{9}$$

$$\begin{aligned}\omega_i &= 1 - \frac{\sum_{k=1}^{x_i} r_k - \Gamma_{\min}}{\Delta\Gamma} = 1 - \frac{2\sum_{k=1}^{x_i} r_k - x_i(x_i+1)}{2x_i(y_i - x_i)} \\ &= \frac{x_i(2y_i - 3x_i - 1) - 2\sum_{k=1}^{x_i} r_k}{2x_i(y_i - x_i)}\end{aligned} \tag{10}$$

In summary, in this section we defined two measures to quantify the consistency of citations in scientific papers, i.e., the tendency to cite papers whose content is related to the paper being prepared. We defined the consistency CI of a paper as the fraction of its references among the most similar papers in the database. CI′ extends CI considering the relative position of papers in the list of the most similar ones. With these two measures we shall quantify the consistency in selecting references for papers in two research areas, namely "graphenes" and "complex networks".

Simulating a systematic search over the citation network

In order to verify the theoretical access frequency to the articles, the dynamics of the citation networks was analyzed using the traditional random walk mechanism, where the next edge is chosen with uniform probability among the adjacent edges. Additionally, we performed a random walk whose probability of transition to a given neighbor is proportional to the number of ingoing edges of the neighbor. Thus, we wished to simulate a walk which tends to choose the most relevant papers for the subject under analysis. To compute the frequency of access in each node of the network we use the matrix W_{ij} and create a Markov chain (Meyn and Tweedie 2005). At each step the network may change its state from the current to another state (or remain in the same state) according to the probability distribution assigned previously. We assume that the transition probabilities are associated with the probability of a reader to follow an edge, which is taken either proportional to the weight of this edge (traditional random walk) or proportional to the in-degree of the neighbor in the end of the edge. Thus, articles with high probability of access at the steady

state (when time tends to infinity) are more likely to be visited in a random walk. Our hypothesis is that if authors perform a systematic search over the citation network, then the articles with the highest access frequency in the steady state are more likely to be consulted (and hence cited). To verify whether this criterion in selecting the papers was obeyed, we calculated the Pearson coefficient (Neter et al. 1996) between the actual number of citations (or downloads) and the expected access frequency from the random walk.

Results and discussion

The main thrust of this paper is to provide formalisms to verify whether possible criteria in selecting references for a paper being prepared are followed. We suggest that two important principles to be followed by authors in a literature survey and in selecting publications to study, and eventually include in the list of references, are: (i) choose strongly related papers (Liu et al. 2009; Huang et al. 2006), which could be done by the similarity of the contents; (ii) perform a systematic search on the citation network involving seminal and strongly related publications. It is true that other criteria could also be used. For instance, the authors could seek the publications considered the most relevant in the topic under study, which could be identified by the number of citations or the reputation of the authors or the journal in which the papers appeared. But using this criterion would inevitably introduce a bias, which we wish to avoid. Besides, such relevant papers are likely to be found in the systematic search anyway. Another possibility is that the authors may need to refer to papers that are only weakly related, but which may for instance provide a methodology from another field used in the research. While this would be certainly justifiable in the selection of references, it is probably of little relevance in terms of the statistics of citation networks, for only a small number of references would be generated if the methodology were not strongly related to the work being performed. Therefore, for the purposes of this paper, we do not need consider the latter criterion. In the following we present results obtained with the overlap and citation networks to show that the two criteria assessed here are not fulfilled.

Consistency index analysis

In order to check whether authors select similar papers to include in their list of references, we defined a consistency index CI which would be 1 if all the references in a paper are the most similar in the dataset and zero when none of the references are among the most similar. The consistency indexes CI and CI′ were computed for the corpus related to the subject of "complex networks" and "graphenes", using either the full text or extracts from the summary and the introduction. The reason why these extracts were used was that the introduction and abstract could be the sections most representative of the topics in a paper. In other words, upon examining only these sections we wished to concentrate on the contents expressing the gist of the paper and its insertion in the research area. Nevertheless, as we shall show the results did not depend on whether the whole manuscripts or only these two sections were analyzed. The texts were represented by complex networks (Antiqueira et al. 2005), whose topology was used to obtain the similarity of contents between two pieces of text (see "Methodology"). The results shown in Tables 1 and 2 point to relatively low consistency indexes for both subjects, regardless of whether the full text or extracts were employed in computing the index. Indeed, in average only 1 out of 4 references appearing in an article was really similar (in the best case). Interestingly, the value of CI′ is

Table 1 Consistency index CI computed for two subjects

Subject and extracts	Cosin	Dice	Jaccard	TSI
Complex networks (abstract)	0.15	0.15	0.14	0.11
Complex networks (abstract and introduction)	0.18	0.18	0.19	0.17
Complex networks (whole paper)	0.22	0.22	0.23	0.20
Graphenes (abstract)	0.13	0.14	0.14	0.11
Graphenes (abstract and introduction)	0.11	0.11	0.10	0.16
Graphenes (whole paper)	0.25	0.21	0.21	0.22

approximately half CI, which means that the citations were distributed regularly in terms of similarity. That is, they are neither concentrated among the most similar nor among the least similar. The conclusions above were not affected by using other methods for calculating the similarity index, such as ignoring the node neighborhood, using the full article or using only articles with many references within the base. In fact, ignoring β_p (see "Methodology") corresponds to calculating the similarity based only on the number of shared words between two texts, without taking into account the topology of the network. These results confirm the hypothesis that other factors are more relevant than the similarity (e.g, it is known that the publication date is relevant, because older articles tend to be forgotten; Redner 1998; Hajra and Sen 2005).

Systematic search over the citation network

The second criterion assessed here was the systematic search which authors should perform on the citation network involving seminal or related papers to the research being conducted. We mimic such search via a random walk search in the citation networks for the two subjects analyzed, namely "complex networks" and "graphenes". To determine the relative frequency of access for each node in the citation network, the transition matrix τ is created by first redefining the citation matrix W_{ij}^{cit} (selecting its greater component) as A_{ij}^{cit} to reflect a reader's browsing on the network. The element A_{ij}^{cit} was 1 if there was an edge $i \to j$ in the citation network W_{ij}^{cit}. If the edge $i \to j$ exists in W_{ij}^{cit} but the edge $j \to i$ does not exist, then one assumes arbitrarily that the weight in $\text{A}_{ij}^{\text{cit}}$ is 0.2, since it is easier for a reader to follow a reference in an article than to find out who cited this article. Hence, we assigned the lower weight to the less probable direction. The matrix τ is a stochastic matrix, derived from $\text{A}_{ij}^{\text{cit}}$. Its definition depends on the type of random walk desired. For the traditional random walk, the probability of transition from node i to node j depends only on the outgoing edges of i, as shown in Eq. 11. In contrast, if the random walk considers the importance of the neighbors to perform the next step, then τ will also depend on the in-degree (k_i) of the neighbors, as shown in Eq. 12.

$$\tau_{ij} = \frac{A_{ij}^{\text{cit}}}{\sum_j A_{ij}} \tag{11}$$

$$\tau_{ij} = \frac{A_{ij}^{\text{cit}} \sum_{k=1} A_{kj}^{\text{cit}}}{\sum_j A_{ij}^{\text{cit}} \sum_k A_{kj}^{\text{cit}}} \tag{12}$$

In both cases, τ measures the probability of a walker visiting a node j after being at node i. We are interested in the stationary, or steady-state, distribution of probabilities of being

Table 2 Consistency index CI' computed for two subjects

Subject and extracts	Cosin	Dice	Jaccard	TSI
Complex networks (abstract)	0.09	0.09	0.09	0.08
Complex networks (abstract and introduction)	0.12	0.12	0.11	0.11
Complex networks (whole paper)	0.15	0.14	0.17	0.15
Graphenes (abstract)	0.06	0.08	0.08	0.07
Graphenes (abstract and introduction)	0.07	0.07	0.06	0.07
Graphenes (whole paper)	0.17	0.14	0.14	0.15

at each node, i.e., one wants to know the probability of being at a given node after an infinite number of steps. This distribution, denoted as π, can be obtained by solving the equation $\tau\ \pi = \pi$. π is the eigenvector associated with the eigenvalue 1 of τ, where the sum of all elements of π is equal to 1 to reflect the sum of probabilities. It is known that for this distribution to be unique, τ needs to be irreducible, i.e. the network must be strongly connected, which happens when every node can reach every other node in the network through a finite path. This property was guaranteed as we selected the greater component of W_{ij}^{cit}.

The Pearson coefficient (Neter et al. 1996) was calculated to infer the correlation between the expected frequency of visits to the nodes (articles) and the actual citations and downloads for the papers in the arXiv repository. The latter information was obtained from the Citebase's site[4], with the number of downloads and number of references to a given article. The results are illustrated in Figs. 4 and 5, which point to low correlations.

The weak correlation probably means that researchers did not consult the references for selecting the papers to cite or to download; in other words, the criterion of a systematic search over the citation network was not fulfilled. Even when performing an efficient search of related works, authors may forget to include papers which were indeed consulted while preparing their manuscript (White and Wang 1997).

Towards virtual scientometry

The results obtained in our analysis clearly indicate, at least in the case of the arXiv database, that two of the most important criteria which should guide citations are not followed in practice. This is worrying because a basic premise of the scientific method, namely that related work should be analyzed to avoid reproduction and warrant originality, is not fulfilled. There is also the side effect that the citation statistics for authors and journals is likely to be strongly biased. Perhaps one should not be surprised with these conclusions. For in a survey sent to psychology journal editors, Cronin (1982) showed that more than 80% of participants believed that authors of papers failed to cite relevant articles, since some citations seemed to be forged to get the attention of editors (Vinkler 1987). There is also the problem of publication date, since there is a tendency to remember recent papers in detriment to older ones (Wang et al. 2009; Gingras et al. 2009). Additional factors known to influence the choice of references have been reported in the literature. For instance, the likelihood to cite a paper may depend on the area of research (Peters and Van Raan 1994; King 1987) (major areas receive more citations than minor areas), on the impact factor of the journal (Vinkler 1987), on the degree of acquaintance among authors

[4] http://www.citebase.org

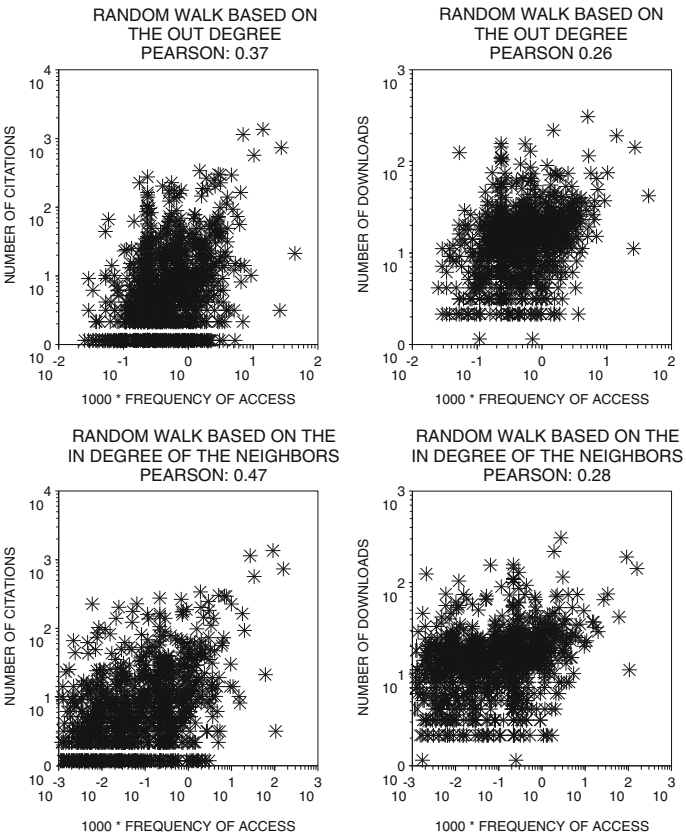


Fig. 4 Scatter plots showing the correlation between the spectral measure computed from the citation network and the real access frequency, for the subject of complex networks. Except for the number of citations compared to the frequency of access of the preferential random walk, all the values for the Pearson coefficient confirm the low correlations

(White 2001), on the type of paper (e.g., survey, regular paper or letters) (MacRoberts and MacRoberts 1997), on the language (Van Raan 2005) and on availability (Lawrence 2001) (e.g., whether the paper is freely online available). Citations are therefore made not only taking into account the intellectual influences from scientific peers, but also non-scientific factors (Bornmann and and Daniel 2008). We believe that means should be devised to complement and/or correct the citation procedures normally adopted. While it would be very difficult to change the established citation habits, the continuing advances in informatics and Internet now allow for innovative approaches to be implemented. Here, we outline a computational approach which could be used to assist authors in identifying relevant references. Furthermore, it could be used in assessing impact of researchers and papers with the concept of virtual citations, i.e., citations which could be artificially established based on e.g. the similarity of contents. This concept is further explained below.

The automated tool for helping authors to identify relevant references could be a simple implementation of the methods reported in the previous sections. More specifically, a manuscript in preparation could be transformed into a network and its content would be

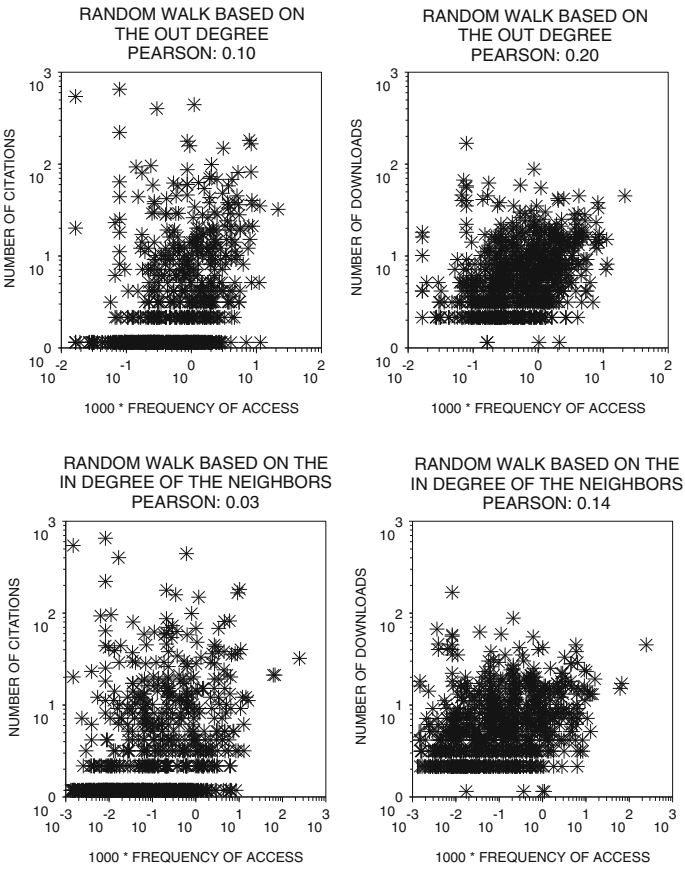


Fig. 5 Scatter plots confirm a weak correlation between the spectral measure computed from the citation network and the real access frequency, for the subject of graphenes

compared to all documents in a given database of articles (e.g. arXiv) using the similarity indexes defined in this work. The output would be a list of related works ranked by decreasing order of overlap/similarity. The author may then check this list and identify potentially relevant works which could otherwise be overlooked. This software tool could also be used to obtain virtual citations for a paper already published, as follows. Given a research area for which a set of papers could be selected, as in the database from the arXiv repository used here, the number of items to be included could be fixed or depend on the similarity of contents among the papers. The virtual citations for a paper would be chosen above a threshold of similarity, which should be defined as to yield an average of citations that is equal to the actual number of references when considering all the papers in the database. Therefore, a paper that is similar to a large number of other papers in the database would receive many citations, as it supposedly deals with a hot topic. Using the citation network obtained from the database, as discussed in the subsection above, another possibility to evaluate the impact of a paper would be to calculate the frequency of visits in a random walk through the network.

The two metrics obtained with the virtual citations and the random walk could be used to compare the impact of an author (or a specific paper) with other authors (or papers), with

the following advantages. (a) No bias from the authors' personal choice exists in selecting the citations, though other types of bias may remain. For example, only papers written in the same language of the manuscript being prepared would be included in the list of references. Moreover, articles outside the database would not be available to be cited. (b) The comparison is straightforward with counterparts in the same field. Therefore, there is no such an effect as a higher impact in a field with higher number of citations per paper than in other fields. It should be emphasized that we are by no means suggesting that the traditional system of citations should be replaced by the virtual citations and the metrics arising from the frequency of visits in the citation network, as proposed here. That could be unreasonable because in addition to similarity the quality of a paper should be considered. However, we believe that the new metrics could be valuable in complementing the assessment of the impact of a given piece of work by reflecting in a comprehensive way its relationship with the literature. Furthermore, authors should be aware that some good practices of literature survey are not being adopted.

Conclusion and further work

In an analysis of a considerably large corpus of 1,800 articles from two research areas, we have found that two important procedures in a literature survey are not followed by authors. It seems that other factors are important, which may not consider the scientific merit. This would be similar to the conclusion in Velho (1986), according to which the number of citations and its impact may depend on the country of the authors, or in Lancaster et al. (1990) where the place of publication also had an effect on the citations. As for the method suggested here to evaluate the contributions from scientists, referred to as virtual scientometry, it is advantageous for being less biased. It is also in line with recent proposals based on a set of metrics to capture the importance of conferences (Martins et al. 2010). In addition, we advocate that software tools should be developed to assist authors in performing a systematic search of the literature. In fact, the communities of medical doctors (Thomas et al. 2004) and software engineers (Barbara 2004) already have well-established methods to conduct surveys (problem formulation, studies selection and data collection are some examples), which are similar to what we propose here—albeit with distinct motivations.

Acknowledgments The authors are grateful to FAPESP (2010/00927-9) and CNPq (Brazil) for the financial support.

References

Aires, R. V. X., Aluísio, S. M., Kuhn, D. C. S., Andreeta, M. L. B., & Oliveira, O. N., Jr. (2000). Combining multiple classifiers to improve part of speech tagging: A case study for Brazilian Portuguese. In *Proceedings of the Brazilian AI symposium.*

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*, 47–97.

Amancio, D. R., Antiqueira, L., Pardo, T. A. S., Costa, L. F., Oliveira, O. N., Jr., & Nunes, M. G. V. (2008). Complex networks analysis of manual and machine translations. *International Journal of Modern Physics C, 19*(4), 583–598.

Amancio, D. R., Nunes, M. G. V., Oliveira, O. N., Jr, Pardo, T. A. S., Antiqueira, L., & Costa, L. F. (2011). Using metrics from complex networks to evaluate machine translation. *Physica A, 390*, 131–142.

Antiqueira, L., Nunes, M. G. V., Oliveira, O. N, Jr., & Costa, L. F. (2005). Modeling texts as complex networks. In *III STIL, Brazilian symposium in information and human language technology*, São Leopoldo, RS, Brazil.

Antiqueira, L., Nunes, M. G. V., Oliveira, O. N., Jr., & Costa, L. F. (2007). Strong correlations between text quality and complex networks features. *Physica A, 373*, 811–820.

Antiqueira, L., Oliveira, O. N., Jr., Costa, L. F., & Nunes, M. G. V. (2009). A complex network approach to text summarization. *Information Sciences, 179*(5), 584–599.

Barabási, A.-L. (2009). Scale-free networks: A decade and beyond. *Science, 24*(325), 412–413.

Barbara, K. (2004). *Procedures for performing systematic reviews*. NICTA Technical Report 0400011T.1.

Börner, K., Marus, J. T., & Goldstone, R. L. (2004). The simultaneous evolution of author and paper networks. *PNAS, 101*(Suppl. 1), 5266–5273.

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? a review of studies on citing behavior. *Journal of Documentation, 64*, 45–80.

Costa, L. F. (2004). What's in a name? *International Journal of Modern Physics C, 15*, 371–379.

Costa, L. F. (2006). On the dynamics of the h-index in complex networks with coexisting communities. arXiv: physics/0609116.

Cotta, C., & Merelo, J. J. (2005). The complex network of evolutionary computation authors: An initial study. arXiv: physics/0507196v2.

Cronin, B. (1982). Norms and functions in citation—The view of journal editors and referees in psychology. *Social Science Information Studies, 2*, 65–78.

De Mey, M. (1982). *The cognitive paradigm*. Chicago: University of Chicago Press.

Ferrer, I., Cancho, R., & Solé, R. V. (2001). The small world of human language. *Proceedings: Biological Sciences/The Royal Society, 268*(1482), 5–2261.

Ferrer, I., Cancho, R., Solé, R. V., & Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E, 69*(5), 1–8.

Gingras, Y., Larivière, V., & Archambault, É. (2009). Literature citations in the internet era. *Science, 323*(5910), 36.

Gross, P. L. K., & Gross, E. M. (1927). College libraries and chemical education. *Science, 66*, 385–389.

Hajra, K. B., & Sen, P. (2005). Aging in citation networks. *Physica A, 346*, 44–48.

Huang, S., Yu, Y., Xue, G.-R., Zhang, B.-Y., Chen, Z., & Ma, W.-Y. (2006). TSSP: Multi-features based reinforcement algorithm to find related papers. *Web Intelligence and Agent Systems, 4*(3), 271–287.

King, J. (1987). A review of bibliometric and other science indicator and their role in research evaluation. *Journal of Information Science, 13*, 261–276.

Lancaster, F. W., Lee, S.-Y. K., & Diluvio, C. (1990). Does the place of publication influence citation behavior? *Scientometrics, 19*(3–4), 239–244.

Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature 411*, 521.

Lilien, G. L. (2008). The ombudsman: Who's at Fawlt at Fawlty Towers? Commentaries on the citation dilemma. *Interfaces, 38*, 123–124.

Liu, Y., Niculescu-Mizil, A., & Gryc, W. (2009). Topic-link LDA: Joint models of topic and author community. In *ICML '09 proceedings of the 26th annual international conference on machine learning*.

MacRoberts, M. H., & MacRoberts, B. R. (1997). Citation content analysis of a botany journal. *Journal of American Society for Information Science, 48*, 5–274.

Martins, W. S., Gonçalves, M. A., Laender, A. H. F., & Ziviani, N. (2010). Assessing the quality of scientific conferences based on bibliographic citations. *Scientometrics, 83*(1), 133–155.

May, K. O. (1967). Abuses of citation indexing. *Science, 19*(156), 890–892.

McClellan, J. E. (2003). Specialist control: The publications committee of the Academie Royal des Sciences. *Transactions of the American Philosophical Society, 93*, 1700–1793.

Meyn, S. P., & Tweedie, R. L. (2005). *Markov chains and stochastic stability*. Cambridge: Cambridge University Press.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). McGraw-Hill/Irwin.

Newman, M. E. J. (2003). The structure and function of complex networks. *Siam Review, 45*(2), 167–256.

Nunes, M. G. V., et al. (1996). O Processo de Construção de um Léxico para o Português do Brasil: Lições Aprendidas e Perspectivas. In *II Encontro para o Processamento Computacional de Portugués Escrito e Falado* (pp. 61–70).

Patrick, D. (1985). A measure of standing of journals in stratified networks. *Journal of the American Society for lnformation Science, 8*(5–6), 341–363.

Peters, H. P. F., & Van Raan, A. F. J. (1994). On determinants of citations scores—A case study in chemical engineering. *Journal of the American Society for Information Science, 27*, 292–306.

Ratnaparki, A. (1997). A maximum entropy part-of-speech tagger. In *Proceedings of the empirical methods in natural language processing conference*, University of Pennsylvania.
Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B, 4*, 131–134.
Shevchuk, R., & Snarskii, A. (2010). Studying the structure of complex networks by the transition to acyclic networks. arXiv: 1010.1864.
Sigman, M., & Cecchi, G. A. (2002). Global organization of the Wordnet lexicon. *Proceedings of the National Academy of Sciences of the United States of America, 99*(3), 7–1742.
Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Boston: Addison-Wesley.
Thomas, J., et al. (2004). Integrating qualitative research with trials in systematic reviews: an example from public health. *British Medical Journal, 328*, 1010–1012.
Van Raan, A. F. J. (2005). For your citations only?. *Scientometrics, 59*, 467–472.
Velho, L. (1986). The meaning of citation in the context of a scientifically peripheral country. *Scientometrics, 9*(1–2), 71–89.
Vinkler, P. (1987). A quasi-quantitative citation model. *Scientometrics, 12*, 47–72.
Wang, M., Yu, G., & Yu, D. (2009). Effect of the age of papers on the preferential attachment in citation networks. *Physica A: Statistical Mechanics and Its Applications, 388*(19), 4273–4276.
White, H. D. (2001). Authors as citers over time. *Journal of the American Society for Information Science and Technology, 52*, 87–108.
White, M. D., & Wang, P. L. (1997). A qualitative study of citing behavior: Contributions criteria, and metalevel documentation concerns. *Library Quarterly, 67*, 122–154.
Wright, M., & Armstrong, J. S. (2008). The ombudsman: Verification of citations: Fawlty towers of knowledge? *Interfaces, 38*, 125–139.