International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:12, No:9, 2018

# Normalizing Scientometric Indicators of Individual Publications Using Local Cluster Detection Methods on Citation Networks

Levente Varga, Dávid Deritei, Mária Ercsey-Ravasz, Răzvan Florian, Zsolt I. Lázár, István Papp, Ferenc Járai-Szabó

***Abstract*—**One of the major shortcomings of widely used scientometric indicators is that different disciplines cannot be compared with each other. The issue of cross-disciplinary normalization has been long discussed, but even the classification of publications into scientific domains poses problems. Structural properties of citation networks offer new possibilities, however, the large size and constant growth of these networks asks for precaution. Here we present a new tool that in order to perform cross-field normalization of scientometric indicators of individual publications relays on the structural properties of citation networks. Due to the large size of the networks, a systematic procedure for identifying scientific domains based on a local community detection algorithm is proposed. The algorithm is tested with different benchmark and real-world networks. Then, by the use of this algorithm, the mechanism of the scientometric indicator normalization process is shown for a few indicators like the citation number, P-index and a local version of the PageRank indicator. The fat-tail trend of the article indicator distribution enables us to successfully perform the indicator normalization process.

***Keywords*—**Citation networks, scientometric indicator, cross-field normalization, local cluster detection.

## I. Introduction

UNBIASED evaluation of scientific quality and impact of an article, researcher or journal is critical to scientific progress. During the last years many efforts have been made to find methods for evaluating research. While the gold standard is to obtain evaluation through peer-review, orchestrating fast and unbiased review is a serious burden for referees, editors and committees. Therefore, simple bibliometric indicators such as the impact factor [1], eigenfactor [2], [3], article influence score [4] or $h$-index [5] are increasingly used. These indicators commonly use citation as an element that indicates a positive review by another author. One of the major shortcomings of the simple bibliometric indicators is that their usage can not assure direct comparison of different disciplines with each other [6], because the publications typically get higher or lower number of citations in different fields. This can result in strong disadvantages for some scientific fields, especially with respect to allocation of public research funding. Accordingly, this is an old topic in scientometrics field, and there have been proposed two major approaches. The cited-side or target normalizations [6], [7] are realized by calculating the citation impact of a publication relative to all publications in the same field. On the contrary, the citing-side or source normalizations [8]–[10] are realized taking into account the referencing behavior of citing publications. It has to be noted, that both approaches may have many pros and cons, and there are a few other ways to realize cross-field normalizations [11].

When evaluating individual publications one can choose to consider the value of the journal, looking for example at its impact factor. However, the distribution of citation numbers of articles appearing in the same journal are strongly asymmetric, so journal indicators cannot give a good prediction about the value of individual articles [12]–[14]. Besides, the impact factor itself has been strongly criticised in the last years [1], [12], [13].

As a result, the most basic way of evaluating individual articles has been to look at the citation number. Citation numbers have been shown to present a fat-tail distribution that depends strongly on the scientific domain (e.g. citations in biology are generally much higher than in mathematics) and the year of publication [7]. A simple cited-side normalization procedure was suggested in [7] where it was shown that citation distributions for publications in different scientific fields can be rescaled to the same universal curve. However, the proposed normalization process supposed to have (1) articles classified into scientific domains, and (2) the average citation number of all articles in each domain. Classification information can be obtained for instance from a priori journal classifications, like scientific categories defined by Journal Citation Report of Web of Science [15]. Then, the average citation number can be arbitrary calculated. However, classifying articles in scientific domains is an ambiguous process, especially as more and more interdisciplinary fields appear [16].

More recently citation networks and their structural properties have been used to define scientometric indicators. In citation networks each node represents a publication and each directed link corresponds to a citation. While the citation

L. Varga, F. Járai-Szabó, D. Deritei, Z. I. Lázár and I. Papp are with Faculty of Physics, Babeş-Bolyai University, 400084 – Cluj-Napoca, Kogălniceanu st. 1, Romania.

M. Ercsey-Ravasz is with Romanian Institute of Science and Technology, Cluj-Napoca, Romania and are with Faculty of Physics, Babeş-Bolyai University, 400084 – Cluj-Napoca, Kogălniceanu st. 1, Romania (e-mail: ercsey.ravasz@phys.ubbcluj.ro).

D. Deritei is with Central European University, Budapest, Hungary.

R. Florian is with Epistemio Systems SRL, Cireşilor st. 29, 400487 Cluj-Napoca, Romania; Epistemio LTD, 145-157 St. John st., London EC1V 4PW, UK.

F. Járai-Szabó is with Faculty of Physics, Babeş-Bolyai University, 400084 – Cluj-Napoca, Kogălniceanu st. 1, Romania (e-mail: jferenc@phys.ubbcluj.ro).

number is simply the in-degree of a node (number of incoming links), more complex measures such as the PageRank [17] or the P-index (h-index of individual articles) [18] are possible to use for evaluating articles. These measures also exhibit strong dependency on scientific domain and time of publication. In this context, several attempts exist for creating classification systems based on large-scale clustering of publications based on citation networks [19]–[22].

Here, considering a further development of the cross-disciplinary target normalization of Radicchi et al. [7] we suggest a procedure that avoids the use of pre-defined domains or any key-word based classification of articles. A systematic procedure for identifying scientific domains based on the structural properties of citation networks is presented. In this procedure the most important technical challenge arises due to the large size of networks. Accordingly, to identify the scientific domain an article belongs to, a local cluster detection algorithm is proposed.

Normalization will be performed in a way that indicators show similar distribution in different domains/clusters. The mechanism of the individual scientometric indicator normalization process is shown for a few article indicators, like the citation number, the P-index [18], and a local version of the PageRank [17] indicator (PageRank calculated on the sub-graph corresponding to a specific domain). Then, using several databases, such as the condensed matter archive at arXiv [23], or the Web of Science database [15] the indicator distributions in different scientific domains are compared to each other and, as a result, a simple article indicator normalization procedure is proposed.

## II. Materials and Methods

### A. Local Cluster Detection Method

In order to perform the normalization of scientometric indicators one needs to identify scientific domains based on the structural properties of citation networks. The betweenness centrality cluster detection algorithm [24] performs well within a variety of networks but it is costly to compute. Many clustering methods are applied regularly in the bibliometric literature [19]–[22] adopted from the existing large number of community detection algorithms [25]. However, citation networks are constantly changing and increasing, so identifying scientific domains as structural clusters of these huge networks needs to be frequently repeated. This is a technical challenge even for cluster detection algorithms that, in general, can deal with large networks.

This constraint puts into focus local cluster detection methods that have low computational costs and can be used in a parallelized fashion [26]. One option to identify the scientific domain an article belongs to would be the use the Local Cluster Detection (LCD) method that consists of a shell spreading outward from a starting vertex. The algorithm is local in the sense that communities can be detected without requiring the partitioning of the entire network. It is based on the idea of Bagrow and Bollt [26]. However, a few improvements have been introduced in order to have a smoother mapping of the studied citation networks.

The proposed algorithm consists of a shell $S_j^l$ spreading outward from a starting vertex $j$. *The shell* is a set of vertices that contains the starting vertex $j$ and another $l$ vertices closest to it according to a specific distance measure. During the shell expansion one node is added in each step: $S_j^{l-1} \subset S_j^l$. The algorithm works by expanding the shell outward from the starting vertex $j$ and comparing in each step the relative number of external edges to a threshold $\beta$. External edges are those ones that have one end inside and the other end outside the shell. On the contrary internal edges are those having both ends inside the shell.

The links of citation networks are inherently directed from the citing article to the cited one. However, in case of cluster detection this character of the network may be neglected because a citation means a kind of common idea relationship of the connected articles. In this sense there is meaningless to talk about the direction of this relationship. Accordingly, the shell expansion process will not respect the orientation of the edges, the algorithm will be applied on undirected citation networks.

*The relative number of external edges* $\kappa_j^l$ of the shell $S_j^l$ is defined by the ratio of its total external degree $K_j^l$ and its total edge number $M_j^l$. The $M_j^l$ includes the total number of edges inside the shell and the number of external links, as well. Accordingly, this measure can be defined as $\kappa_j^l = \frac{K_j^l}{M_j^l}$. Example values of these shell-parameters for a small sample network are shown in Fig. 1.

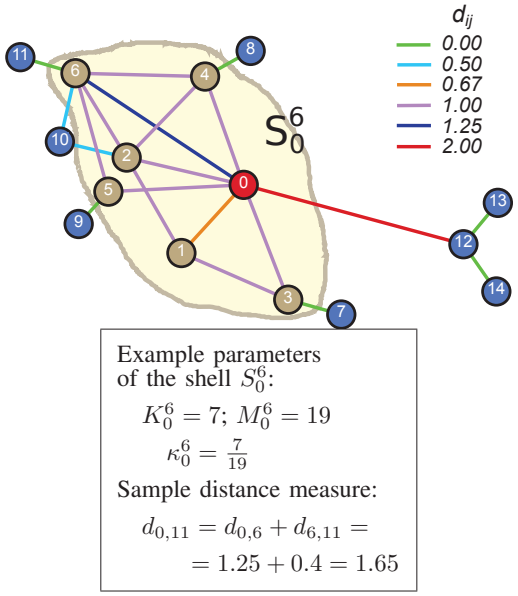


Fig. 1 The sketch of the shell-growing mechanism shown on a small sample network. The red node is the starting node of the shell (with index 0), beige nodes are its direct neighbors. Nodes are numbered according to their distance from the starting node. Shell $S_0^6$ contains all nodes with index up to 6. The characteristic quantities for this shell ($j = 0, l = 6$): the values of the total outgoing degree $K_j^l$, total edge number $M_j^l$, and the relative number of external edges $\kappa_j^l$ are listed near the figure, together with a sample distance measure calculation

Here we note, that the original shell-growth stopping condition of Bagrow and Bollt [26] that counted for the change

in the total external degree $\Delta K_j^l = \frac{K_j^l}{K_j^{l-1}}$ has been replaced in order to adopt the method for citation networks. In case of these networks the $\Delta K_j^l$ quantity, except for a very short transition region, is fluctuating around 1. Accordingly, this cannot be used to compare to a threshold. At the same time, a clear trend in the relative number of external edges has been observed (even with a strong minimum in case of some networks) which makes it comparable to predefined thresholds.

The other important ingredient of the LCD algorithm is the distance measure. In order to find a proper distance measure we have to keep in mind that our task is to detect clusters in a citation network starting from a certain article. Accordingly, the distance has to deliver information about how strongly neighboring nodes are connected to each-other. In this sense one possible option would be the use of the Edge Clustering Coefficient (ECC) $C_{i,j}$. This was first introduced by Radicchi [27] and is defined as the total number of triangles $z_{i,j}$ an edge connecting the nodes $i$ and $j$ belongs to, divided by the number of triangles that might potentially include it (given by the degrees of the two adjacent nodes $k_i$ and $k_j$). In order to handle leaf nodes, when the number of triangles is zero, an additional constant has been added to the expression. Accordingly, ECC is computed as

$$C_{i,j} = \frac{z_{i,j}+1}{\min[(k_i-1),(k_j-1)]}\,. \tag{1}$$

Then, the distance measure used in the proposed algorithm may be defined as the reciprocal value of ECC.

$$d_{i,j} = \frac{1}{C_{i,j}}\,. \tag{2}$$

We trust this distance measure, because it has already been successfully used in a network clustering algorithm based on graph Voronoi diagrams [28]. This measure is defined for neighboring nodes $i$ and $j$, therefore the distance between any pair of nodes needs to be calculated as the length of shortest path in the network. Then, in each shell-growth step $l$ only one single node having the shortest distance to the initial node is added to the expanding shell $S_j^l$. In case of distance-equality the nodes are randomly sorted.

In order to illustrate the shell-growing mechanism, in Fig. 1 the shell $S_0^6$ is shown on a sample network. The node labels are ordered according to their distance to the initial node 0. As a result, in this simple example the label of a node shows us the shell-growth step number $l$ at which it was added or it will be added later to the shell.

First, we tested the LCD algorithm on a small benchmark network of $N = 200$ nodes and $M = 4585$ edges. The network is generated by the benchmark software framework provided by Lancichinetti et al. [29]. For this first trial, the mixing parameter, which sets the rate of edges inside and between clusters is selected to be a small value of $\mu = 0.1$. The network layout shown in Fig. 2a was created by the ForceAtlas2 algorithm of the *Gephi* network visualization software [30].

The panels b-f of Fig. 2 show the shell-growth process for different shell sizes of 25, 50, 75, 90 and 175 nodes, respectively. As it is immediately observable, after the whole cluster is detected at growth step 90, the shell-growing process continues by including nodes from multiple other clusters.

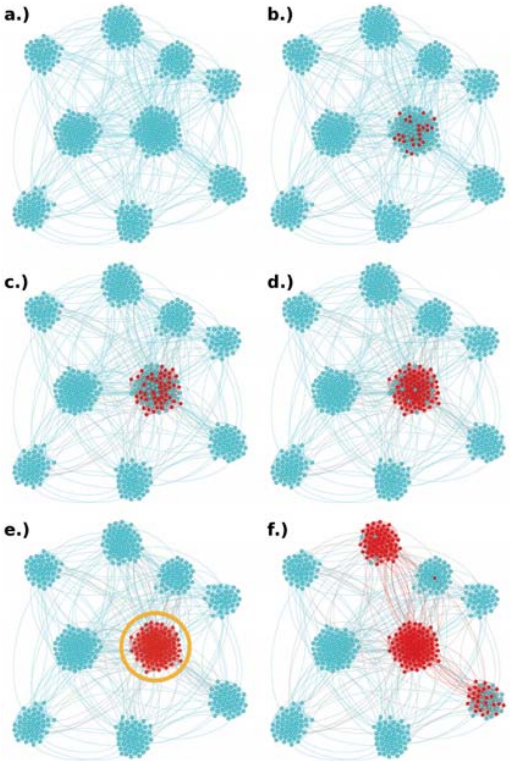


Fig. 2 The shell-growth process on the small benchmark network with $N = 200$, $M = 4585$ and $\mu = 0.1$. The different panels show the different stage of the shells containing (a) 0, (b) 25, (c) 50, (d) 75, (e) 90, (f) 175 nodes

In order to find the growth step at which the shell-growing process has to stop, the relative number of external edges $\kappa_j^l$ of the shell is calculated. The results are presented in Fig. 3, where different curves represent results for shells started from different initial nodes of the same cluster. Results show for each curve a first sharp minimum at the shell-growth step $l = 90$, which means, that in case of each starting node $l$, the same cluster has been constructed.

Accordingly, in case of benchmark networks a well defined shell-growth stopping condition is found. Namely, in order to detect the cluster to which a node $j$ belongs, the shell-growth process has to be started with the respective node and it has to be stopped when the relative number of external edges of the shell $\kappa_j^l$ reaches its first local minimum.

For further testing the LCD clustering of the Political blogosphere network [31] is also constructed and it is presented in Fig. 4. The network contains $N = 1490$ nodes and $M = 19025$ edges. Here, two different starting nodes $j$ are selected randomly from the two clusters showed separately by the ForceAtlas2 layout of the *Gephi* network visualization software [30]. On the right panel of the figure, the obtained shells are colored by red and blue, respectively. According to LCD the yellow nodes belongs to both clusters, and the cyan nodes belongs to none of these two clusters. The $\kappa_j^l$ curve is shown in the left panel of Fig. 4. In order to remove

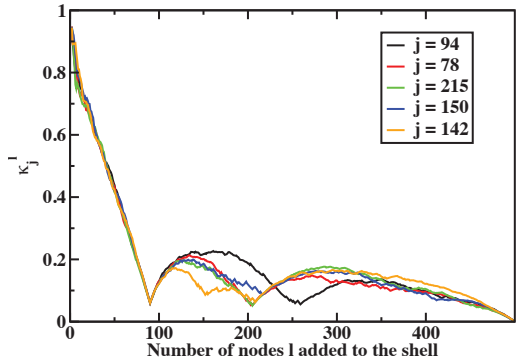


Fig. 3 The relative number of external edges of the shell $\kappa_j^l$ as a function of the number of nodes $l$ added to the shell. The different curves represent shells started from different nodes ($j$) of the same cluster. Results are presented for the sample network shown in Fig. 2

small fluctuations the $\kappa_j^l$ values are averaged by a moving window of size 10. The shell-growth stopping conditions are marked by vertical dashed lines on the $\kappa_j^l$ graph. The exact position of these two lines is shown on the magnified inset graphs, as well. By the LCD algorithm two large clusters are successfully detected from two starting nodes that belong to different clusters. We see that even in this case, the $\kappa_j^l$ curves have a well-detectable minima which can be used as a stopping condition for the shell-growth process.

The LCD method has been tested on many other real-world networks, as well. The shell-growth stopping condition is not so straightforward in all cases, especially when clusters are not se well separated (network with lower modularity). In contrast to the Political blogosphere network, there are some cases, where the $\kappa_j^l$ curve has no well-detectable minimum. Accordingly, in such conditions, an arbitrary shell-growth stopping condition has to be applied. Namely, we have to choose a threshold value $\beta$ for the relative number of external degrees. This is done by setting the threshold value low enough to get clusters with a number of nodes that can be treated statistically. As we will show later, the results are not too sensitive on the selection of threshold value.

### B. Calculation of Article Indicators

A few selected article indicators, like the citation number, the P-index (h-index of individual articles) [18], and a local version of the PageRank [17] indicator (PageRank calculated on the sub-graph corresponding to a specific domain) will be considered later for normalization. Here, these indicators are shortly described.

The citation of one article by another is characteristic in science. Therefore, the number of citations of an article reflects its impact in the scientific community [32]. Accordingly, *the citation number* may be considered as the most simple bibliometric indicator of a published article. In the citation network of articles this is equal with the in-degree of nodes.

*The PageRank* [17] on a graph is a probability distribution that represents the likelihood that a random walker visiting the graph through its edges will arrive at any particular vertex. Unlike for webpage networks, in case of scientific citation networks due to the time ordering of the link creation there is no possibility for loop formation which may be considered as "rank sink". Accordingly, it is enough to operate with a simplified version of the PageRank [17] which is equivalent to the eigenvector centrality where this "rank sink" loops are not treated. This simplified PageRank of an article is calculated through iterative steps. Initially, each vertex $u$ has a PageRank value of $R(u) = 1$. Then, in each iteration step the new PageRank of each vertex $u$ is calculated as

$$R(u) = \sum_{v \in N_u} \frac{R(v)}{\deg(v)},$$

where $N_u$ represents the set of neighboring vertices of $u$, while $\deg(v)$ denotes the degree of vertex $v$. These iteration steps will be continued until convergence is reached. The *local PageRank* refers to PageRank values calculated only on a local cluster of the graph.

The Hirsch index (h-index) has been defined for the evaluation of scientists or scientific groups [5]. The h-index of a scientist is the maximal number $h$ such that he/she has at least $h$ publications, which have at least $h$ citations each. This may be adopted to individual publications, as well. *The P-index of a publication* is the maximal number $P$ such that the publication is cited by at least $P$ publications, which have at least $P$ citations each [18].

## III. Scientometric Indicator Normalization

Citation practices vary between different fields of science [4], [14], [33]. The most simple way to characterize the average citing behavior in a field is through the statistics of different article indicators. The probability distribution of an article indicator shows us the occurrence probability of articles with certain article indicator values. Accordingly, these probability distributions may be characteristic to the citation behavior in different scientific domains. The scientometric distribution functions have been intensively studied in the last decades [7], [34]. The field variation of these distributions affect the cross-disciplinary evaluation of individual publications, journals and researchers, as well. Accordingly, there is an intensive effort to obtain field-independent indicators [35], [36] or to normalize the existing article indicator distributions across different scientific fields [7].

First, let us study the in-degree (citation) distributions on clusters of two different modular benchmark networks [29] *bmn-1* and *bmn-2*. These networks are composed by 50 000 nodes each, the mixing parameter of both is $\mu = 0.1$. These networks differ only on the average degree of nodes. For bmn-1 and bmn-2 networks, the average out-degree is 30 and 120, respectively. From both networks three different clusters are constructed using the LCD algorithm starting from randomly selected nodes. The in-degree distribution of each cluster is represented in panel A of Fig. 5.

It is immediately observable, that in-degree $n_i$, or citation distribution functions of different clusters on the same network falls to a single straight curve on the log-log plot. Accordingly, the distribution function can be described by the

$$p(n_i) = An_i^{\alpha} \tag{3}$$

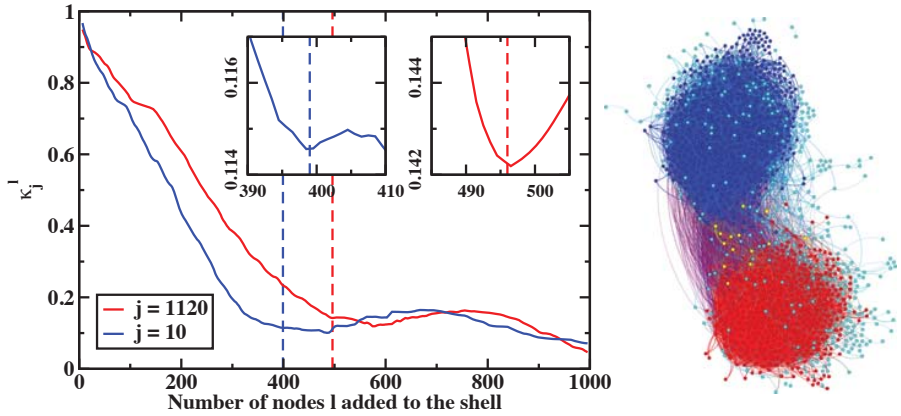


Fig. 4 LCD clustering of the Political blogosphere network [31]. On the left panel the $\kappa_j^l$ graphs are shown on which the vertical dashed lines represent the $l$ value at which the shell-growth process is stopped. The magnified inset graphs show the exact position of these vertical lines. On the right side the detected clusters are colored

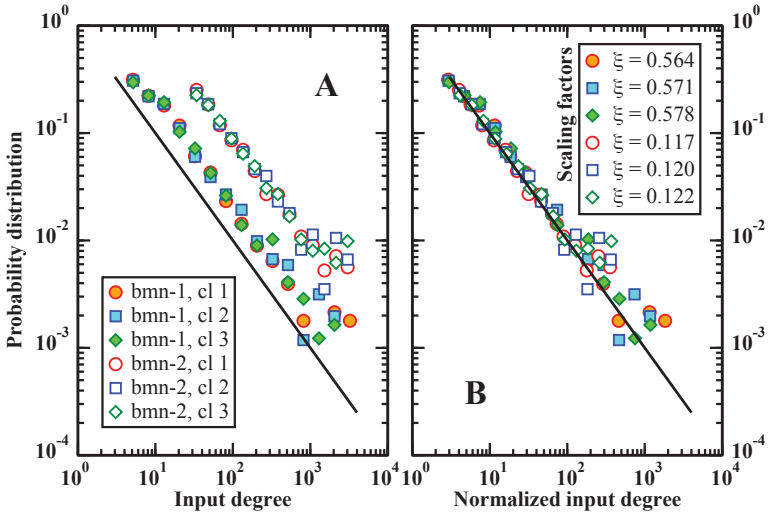


Fig. 5 In-degree distributions (A) and normalized in-degree distributions (B) on different clusters of two benchmark networks

power-law, where $A$ is a normalizing constant. Here, the exponent was found to be $\alpha = -1.0$. For a visual check, the $p(n_i) = n_i^{-1}$ curve is drawn by solid line in Fig. 5. The distribution functions of clusters taken from different networks have the same exponent $\alpha$, and they differ only by the value of normalizing constant $A$, which has values about $A = 1.73 - 1.78$ for clusters of bmn-1, and $A = 8.2 - 8.6$ for clusters of bmn-2. The numerical values of fitting parameters are summarized in columns 2-7 of Table I. The $R^2 > 0.96$ (correlation coefficient) values indicates the goodness of the fits. The average out-degree of the represented clusters varies between $\langle d \rangle = 28 - 29$ for bmn-1, and $\langle d \rangle = 124 - 136$ for bmn-2. Thus, the normalizing constant looks to be somehow connected to the out-degree of the clusters. This behaviour has also been tested in case of other benchmark networks having different mixing parameter values $\mu$.

This finding encourages us to assume that the citation behavior (i.e. the average number of citation in a scientific field) is influencing the distribution of citations only by a scaling factor. Accordingly, the normalization of article indicators that are based on the citation number may be realized using this factor. In case of our benchmark networks one option would be to normalize each distribution to $A = 1$ (solid line in Fig. 5). Accordingly, each in-degree has to be multiplied by a scaling factor $\xi = A^{1/\alpha}$, as follows:

$$p(n_i) = An_i^{\alpha} = (A^{1/\alpha} n_i)^{\alpha} = (\xi n_i)^{\alpha}. \qquad (4)$$

The normalized distributions and citation scaling factors are shown in panel B of Fig. 5. Here one single power-law function has been fitted to all normalized distribution data. The numerical values of fitting parameters are presented in column 8 (bmn-all) of Table I. The high value for the $R^2$ correlation coefficient indicates that we have succeeded the normalization procedure.

In the following, these assumptions on article indicator scaling are tested on different real citation networks. First, let us show article indicator distributions on different clusters of the collaboration network of scientists posting preprints on the condensed matter archive at the arXiv database (CONDMAT). This is based on preprints between January 1, 1995 and March

TABLE I
FITTING PARAMETERS FOR THE DATA IN FIG. 5, FOR DIFFERENT BENCHMARK DATA, CONSIDERING (3). THE $R^2$ CORRELATION COEFFICIENT VALUES ARE ALSO INDICATED FOR EACH FIT

| | bmn-1, cl 1 | bmn-1, cl 2 | bmn-1, cl 3 | bmn-2, cl 1 | bmn-2, cl 2 | bmn-2, cl 3 | bmn-all |
|---|---|---|---|---|---|---|---|
| $A$ | 1.77 | 1.75 | 1.73 | 8.58 | 8.36 | 8.21 | 1.00 |
| $\alpha$ | -1.0 | -1.0 | -1.0 | -1.0 | -1.0 | -1.0 | -1.0 |
| $R^2$ | 0.980 | 0.977 | 0.969 | 0.998 | 0.996 | 0.995 | 0.983 |

31, 2005 [23] and it contains 39 576 nodes and 175 692 links. The citation number, the local PageRank, and the P-index article indicator distributions are presented in Fig. 6. These distributions are calculated for four different clusters of the CONDMAT network detected by the LCD algorithm starting from randomly selected nodes. We have to note that for this network no automatic shell-growth stopping condition (i.e. detecting minima of some relevant parameter) has been found. Accordingly, an arbitrary selected threshold value $\beta = 0.4$ has been applied. The effect of this $\beta$ parameter on article indicator distributions will be described later.

Looking at panel A of Fig. 6, it can be easily recognized that the in-degree or citation distributions fall almost on the same curve as in case of clusters from one single benchmark network. Accordingly, in case of this network no scientific indicator normalization is needed. This conclusion can be understood, if we take into account that the CONDMAT network is a citation network of the Condensed Matter scientific field, where one expects sub-domains with quite similar citation practices. In order to check this, we have counted the average citation numbers in the studied clusters and it was found that these values are in a narrow interval $\langle k \rangle = 5 - 8$.

It is well known that citations have fat-tailed distributions with tails often described in terms of power-laws [37], [38]. In agreement with this, the tail of citation distribution of different clusters of the CONDMAT network looks to be a power-law for a range of two orders of magnitude. For visual comparison the power-law trend with exponent $-2$ is shown with solid line in Fig. 6A. This power-law trend has been recently criticized showing that the power-law hypothesis is rejected for around half of the Scopus fields of science [39]. We do not want to contribute to this debate, because in case of our study the domains are detected by the LCD algorithm and no pre-defined scientific sub-domain structures are used. This may alter our citation distributions from those found by other previous studies.

Up to our vision, this fat-tail trend may be useful in citation normalization that may be realized by scaling the citations in such way that these tail-distributions fall onto the same curve. Such article indicator scalings will be illustrated below in case of another citation network.

Beside the citation distributions, we have looked at the local PageRank distributions in panel B and the P-index distributions in panel C of Fig. 6. Similarly to the case of the citation indicator, all distributions constructed on different clusters fall almost onto the same curve, so no normalization is needed in this particular case. However, the power-law-type fat-tail trend in case of P-index is not detectable and for this indicator we have to find another way to realize the cross-disciplinary normalization. The proposed solution will be discussed below in case of another citation network in which distributions in different clusters do not fall onto the same curve, therefore normalization of the P-index will be needed.

Further, we have used a citation network from the Web of Science database [15] that contains 771 914 nodes (publications) connected by 7 779 703 links (references). It has to be noted that the present network is a complete subgraph of the Web of Science network, and it contains a significant number of nodes and links which enables us to use our statistical approach.

The WoS-net has been clusterized using the Louvain algorithm [40], which found a number of 59 large clusters. In this network, different scientific fields are present which are not necessarily sub-fields of a single scientific domain as in case of our CONDMAT network studies. Accordingly, here different article indicator distribution scalings are expected. This presumption is further supported by the wider distribution of average citation numbers in these clusters. The average out-degree values fall in the interval $\langle k \rangle = 2 - 20$ in a few studied clusters of the network.

The article indicator distributions are represented for four different clusters detected by the LCD algorithm starting from randomly selected nodes in the left panels of Fig. 7. Again, as in case of the CONDMAT network the arbitrarily selected $\beta = 0.4$ threshold value has been used.

As in case of our previous studies, the citation (in-degree) distributions in panel A of Fig. 7 have a fat-tailed distribution, and in agreement with our expectation the distributions of different clusters differ considerably. Here, for simplicity, the fat-tail of the citation distributions will be fitted again by the power-law distribution function (3). The exponent is found to be close to $\alpha = -1.0$. The normalization to $A = 1$ (solid line on Fig. 7B) can be realized following the same procedure described previously in case of benchmark networks. The normalized citation distributions and the corresponding scaling factors $\xi$ are represented in Fig. 7B. The wider range of scaling factors $\xi = 0.6 - 2.05$ shows us that in case of WoS-net it make sense to scale the citation numbers in order to obtain cross-disciplinary normalization of this article indicator.

If one looks to the local PageRank distributions presented in panel C of Fig. 7, the same conclusions can be drawn. The tail of the distribution function can be fitted by a power-law and the normalization procedure can be realized. Here, the PageRank values are scaled to the distribution function $p(n_i) = 10^{-7} n_i^{-1.4}$. The result of this scaling and the scaling factors are shown in panel D of the same figure. Due to the different nature of the local PageRank article indicator (i.e. it is calculated only on a local cluster of the citation

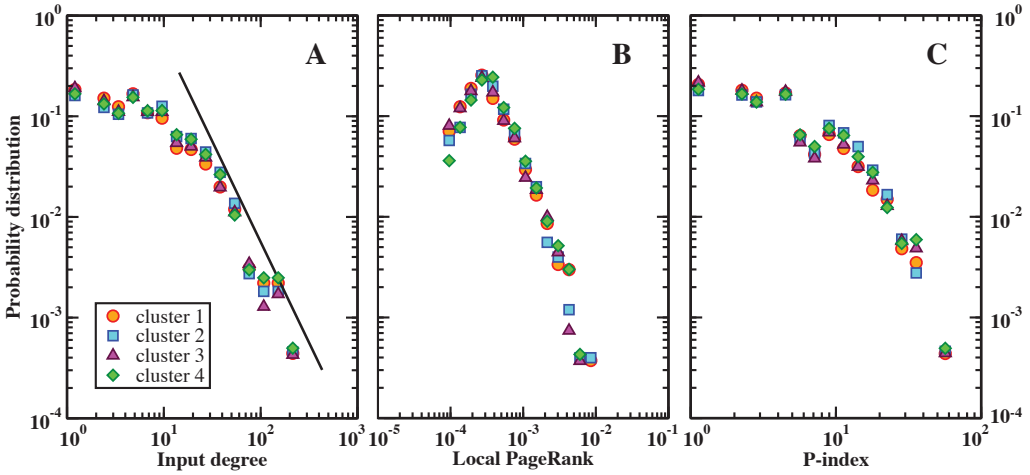


Fig. 6 In-degree (A), local PageRank (B) and P-index (C) distributions on different clusters of the CONDMAT network [23]

network) the obtained scaling factors differ significantly from those obtained in case of citation.

Finally, the P-index distributions are studied, as well. Here, the power-law trend is not so evident and as a consequence, we cannot realize the normalization according to distribution tails. However, a method used for normalizing the H-index [7] can be successfully adopted. We propose, to realize an approximate normalization based on the already studied citation indicator. The P-index of an article is actually calculated by ranking in descending order by the number of citations the articles that cite the article in focus. In this list the total number of items is equal to the citation number of the article. The same procedure can be realized taking into account the citation numbers scaled by a factor $\xi$. Here, our main presumption is that the majority of citing articles belong to the same scientific field as the article in focus and accordingly, the same scaling factor can be used for all of them. Then, in order to get the total number of items in the list equal to the scaled citation number of the article, the ranking index has to be scaled by $\xi$. Therefore, the P-index of the article will be scaled approximately by the same factor $\xi$. P-indices scaled by the factor obtained for citations is shown in Fig. 7F. In contrast to the original, unscaled values shown in panel E of Fig. 7, the scaled indicator distribution functions are much closer to each other. Accordingly, using the same scaling factors as in case of citations, the cross disciplinary normalization of P-indices can be approximated. Moreover, we think that this method can be generalized for other scientometric indicators (like journal impact factor) that are based on citation numbers, as well. Cross-disciplinary normalization can be obtained for such indicators if their calculation would be realized with already scaled citation numbers.

In the presented case-studies, we used the LCD cluster detection algorithm with an arbitrary selected shell-growth stopping threshold $\beta = 0.4$. In order to study the robustness of the proposed cross-disciplinary normalization procedure regarding to this threshold we realized another study. The same starting node as in case of cluster 3 of the previously studied WoS-net citation network has been selected and three local clusters have been constructed by LCD using thresholds $\beta = 0.3, 0.4$, and $0.5$. The obtained citation distributions are presented in panel A of Fig. 8. From the figure it is obvious, that the distribution functions fall to the same curve, except the $\beta = 0.5$ case, where major fluctuations can be detected. This is attributed to the fact that in case of large threshold values the obtained clusters are small and the small number of citation data can lead to fluctuations in distribution functions. Accordingly, we can state, that if we deal with clusters large enough to be treated statistically, the resulted citation distributions do not depend strongly on the $\beta$ parameter of the LCD algorithm. On the contrary, local PageRank indicator distributions represented on panel B of Fig. 8 show a stronger dependence on the parameter $\beta$. This trend can be understood, if we take into account that the citation data is calculated on the whole network, while the local PageRank values by definition are calculated only on the cluster in focus.

## IV. Conclusion

A cross-disciplinary scientometric indicator normalization procedure was presented based on the structural properties of citation networks. For the identification of scientific domains a local cluster detection algorithm has been used which is capable to treat large citation networks. The algorithm was used to detect the scientific field of the article in focus. Then, the article indicator distribution function has been constructed in this sub-graph, and the fat-tail trend of the distribution was used for the article indicator normalization.

The proposed procedure has been tested on benchmark graphs and on a complete subgraph of the Web of Science citation network. We found that the procedure is capable to successfully normalize globally calculated simple article indicators. In case of more complex indicators the cross-disciplinary normalization has been obtained if the calculation of these article indicators was realized based on some already scaled basic indicators. In this sense the normalization of the P-index has been shown based on the citation number basic indicator. Based on these results we believe, that the method used for P-index may be generalized

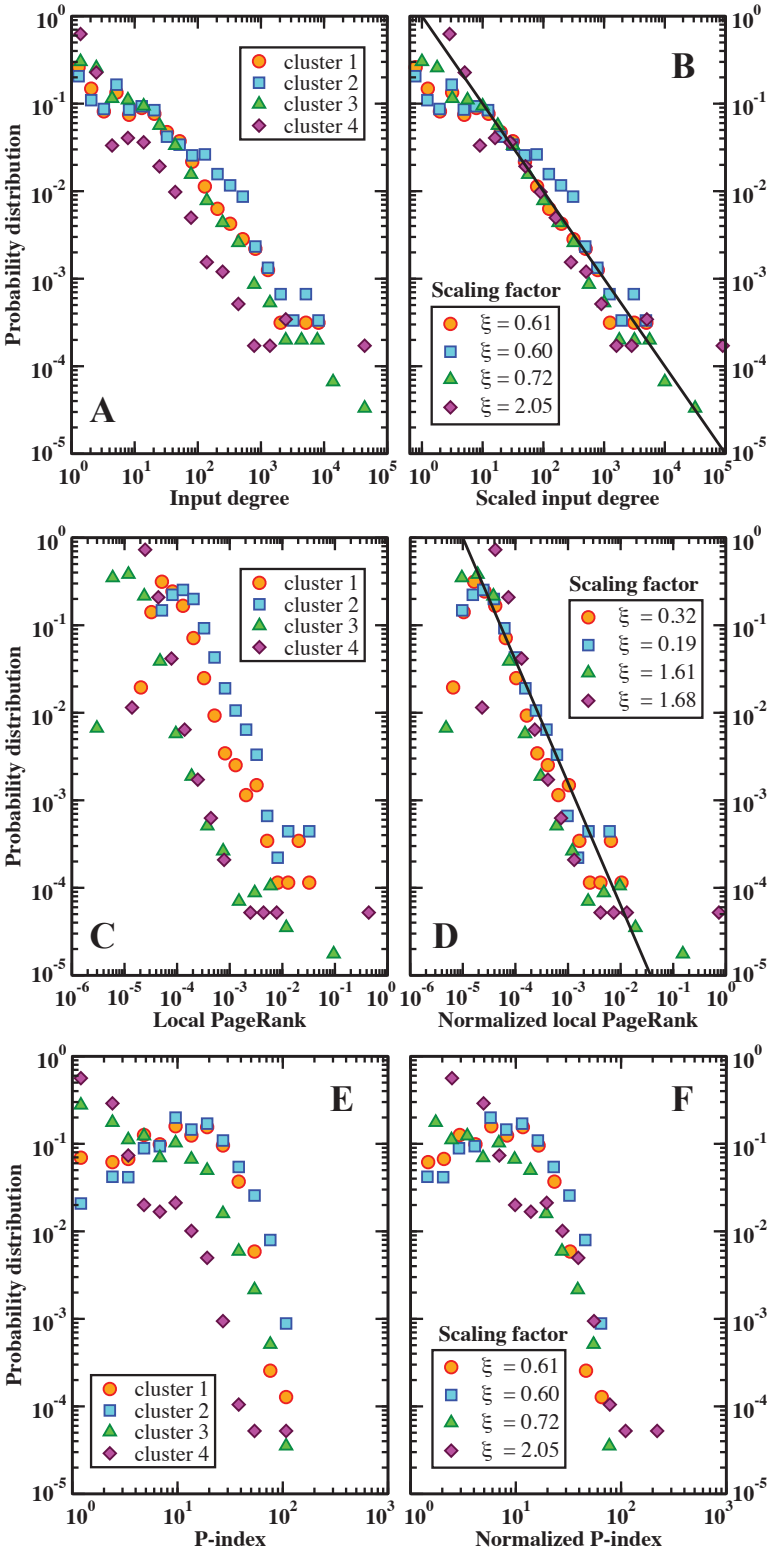


Fig. 7 In-degree (A), local PageRank (C) and P-index (E) distributions on different clusters of a citation network extracted from the Web of Science (WoS-net) [15]. The normalized article indicators and scaling factors are presented in the right-side panels (B, D, F) of figure

even for impact factor normalization, as well, because it is a measure reflecting the yearly average number of citations to recent articles published in that journal.

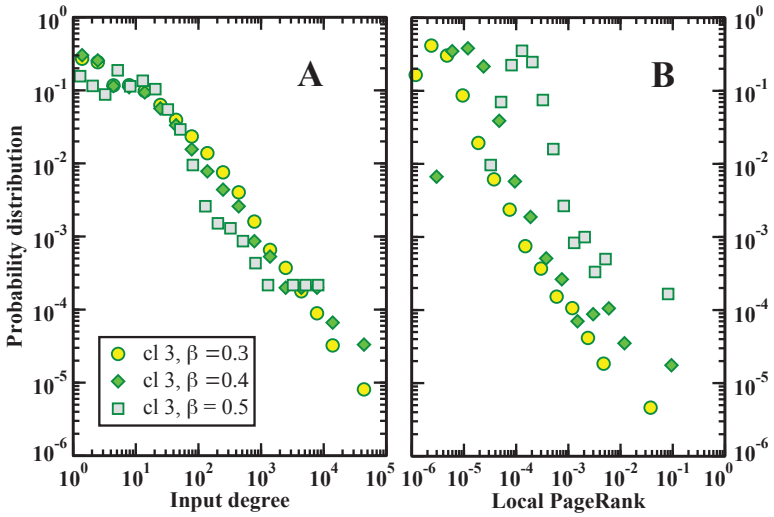


Fig. 8 In-degree (A) and local PageRank (B) distributions of the cluster 3 of WoS-net used in Fig. 7 for different shell-growth stopping thresholds $\beta$

## Acknowledgment

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS/CCCDI–UEFISCDI, project number PN–III–P2–2.1–BG–2016–0252, within PNCDI III.

## References

[1] Garfield, E. (1998). The Impact Factor and Using It Correctly. Der Unfallchirurg, 101(6), 413–414.
[2] Bergstrom, C. T., West, J. D., Wiseman, M. A. (2008). The eigenfactor metrics. Journal of Neuroscience, 28(45), 11433–11434.
[3] Davis P. M. (2008). Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts. JASIST, 59(13), 2186–2188.
[4] Bollen, J., Rodriguez, M. A., Van de Sompel, H. (2006). Journal status. Scientometrics, 69(3), 669–687.
[5] Hirsch, J. E. (2005). An index to quantify an inidividual's scientific research output. PNAS, 102(46), 16569–16572.
[6] Schubert, A., Braun, T. (1996). Cross-field normalization of scientometric indicators. Scientometrics, 36(3), 311–324.
[7] Radicchi, F., Fortunato, S., Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. PNAS, 105(45), 17268–17272.
[8] Waltman, L., van Eck, N. J. (2013). Source normalized indicators of citation impact: an overview of different approaches and an empirical comparison, Scientometrics, 96, 699, https://doi.org/10.1007/s11192-012-0913-4.
[9] Bouyssou, D., Marchant, T. (2016). Ranking authors using fractional counting of citations: An axiomatic approach. Journal of Informetrics, 10(1), 183–199, https://doi.org/10.1016/j.joi.2015.12.006.
[10] Zitt, M., Small, H. (2008) Modifying the journal impact factor by fractional citation weighting: The audience factor. J. Am. Soc. Inf. Sci., 59(11), 1856–1860, http://dx.doi.org/10.1002/asi.20880.
[11] Kostoff, R. N. (1997) Citation analysis cross-field normalization: a new paradigm. Scientometrics., 39(3), 225-230.
[12] Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. BMJ, 314(7079), 498–502.
[13] Opthof, T. (1997). Sense and nonsense about the impact factor. Cardiovascular Research, 33(1), 1–7.
[14] Bornmann, L., Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. Journal of Documentation, 64(1), 45–80.
[15] Web of Science. http://www.webofknowledge.com, Accessed on 07/05/2018.
[16] Leydesdorff, L., Wagner, C. S., Bornmann, L. (2017). Betweenness and diversity in journal citation networks as measures of interdisciplinarity A tribute to Eugene Garfield. Scientometrics. https://doi.org/10.1007/s11192-017-2528-2.
[17] Page, L., Brin, S., Motwani, R., Winograd, T. (1999). The PageRank citation ranking: Bringing order to the Web, Technical Report 1999-66, Stanford InfoLab, November 1999. http://ilpubs.stanford.edu:8090/422/. Accesed 21 August 2013.
[18] Papp, I., Ercsey-Ravasz, M., Deritei, D., Sumi, R., Járai-Szabó, F., Florian, R. V., Cabuz, A. I., Lázár, Zs.I. (2013). The P-Index: Hirsch Index of Individual Publications. Proceedings of ISSI, 2013, 2086–2088.
[19] Waltman, L., van Eck, N. J. (2012), A new methodology for constructing a publicationlevel classification system of science, J. Am. Soc. Inf. Sci. Technol., 63, 2378–2392.
[20] Ruiz-Castilloa, J., Waltman, L. (2015), Field-normalized citation impact indicators using algorithmically constructed classification systems of science, Journal of Informetrics, 9, 102–117.
[21] Šubelj, L., van Eck, N. J., Waltman, L. (2016), Clustering Scientific Publications Based on Citation Relations: A Systematic Comparison of Different Methods. PLOS ONE, 11, e0154404.
[22] van Eck, N. J., Waltman, L. (2017) Citation-based clustering of publications using CitNetExplorer and VOSviewer. Scientometrics, 111, 1053-1070.
[23] Newman, M. E. J. (2001). The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA, 98(2), 404–409.
[24] Newman, M. E. J., Girvan, M. (2004). Finding and evaluating community structure in networks. Phys. Rev. E, 69, 026113.
[25] Castellano, C., Fortunato, S., Loreto, V. (2009). Statistical physics of social dynamics. Rev. Mod. Phys., 81, 591.
[26] Bagrow, J. P., Bollt, E. M. (2005). Local method for detecting communities. Phys. Rev. E, 72(4), 046108.
[27] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D. (2004). Defining and identifying communities in networks. PNAS, 101, 2658–2663.
[28] Deritei, D., Lazar, Zs.I., Papp I., Jarai-Szabo, F., Sumi, R., Varga, L., Regan, E., Ercsey-Ravasz, M. (2014). Community detection by graph Voronoi diagrams. New Journal of Physics, 16, 063007.
[29] Lancichinetti, A., Fortunato, S., Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. Phys. Rev. E, 78, 046110.
[30] Bastian, M., Heymann, S., Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. ICWSM, 8, 361–362.
[31] Adamic, L. A., Glance, N. (2005). The political blogosphere and the 2004 US Election. Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem.
[32] Durieux, V., Gevenois, P. A. (2010). Bibliometric indicators: quality measurements of scientific publication. Radiology, 255(2), 342–351.
[33] Hargens, L. L. (2000). Using the literature: reference networks, reference contexts, and the social structure of scholarship. Am Sociol Rev, 65, 846–865.

[34] Van Raan, A. F. J. (2006). Statistical Properties of Bibliometric Indicators: Research Group Indicator Distributions and Correlations. J. Am. Soc. Inf. Sci. Tec., 57, 408–430.
[35] Hutchins, B. I., Yuan, X., Anderson, J. M., Santangelo, G. M. (2016). Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level. PLOS Biology., 14, e1002541. https://doi.org/10.1371/journal.pbio.1002541.
[36] Gonzalez-Betancor, S. M., Dorta-Gonzalez, P. (2017). An indicator of the impact of journals based on the percentage of their highly cited publications. Online Information Review, 41, 398–411.
[37] Tsallis, C., De Albuquerque, M. P. (2000). Are citations of scientific papers a case of nonextensivity? Eur. Phys. J. B, 13, 777–780.
[38] Lehmann, S., Lautrup, B., Jackson, A. D. (2003). Citation networks in high energy physics. Phys. Rev. E, 68, 026113.
[39] Brzezinski, M. (2015). Power laws in citation distributions: evidence from Scopus. Scientometrics, 103, 213–228.
[40] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., Lefebvre E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech., P10008.