Scientometrics
DOI 10.1007/s11192-012-0762-1

The *d*-index: Discovering dependences among scientific collaborators from their bibliographic data records

Luigi Di Caro · Mario Cataldi · Claudio Schifanella

Received: 1 August 2011
© Akadémiai Kiadó, Budapest, Hungary 2012

Abstract The evaluation of the work of a researcher and its impact on the research community has been deeply studied in literature through the definition of several measures, first among all the *h*-index and its variations. Although these measures represent valuable tools for analyzing researchers' outputs, they usually assume the co-authorship to be a proportional collaboration between the parts, missing out their relationships and the relative scientific influences. In this work, we propose the *d-index*, a novel measure that estimates the dependence degree between authors on their research environment along their entire scientific publication history. We also present a web application that implements these ideas and provides a number of visualization tools for analyzing and comparing scientific dependences among all the scientists in the DBLP bibliographic database. Finally, relying on this web environment, we present case and user studies that highlight both the validity and the reliability of the proposed evaluation measure.

Keywords Evaluation metrics · Social networks · Research evaluation · Collaboration analysis · Data visualization

Introduction

The evaluation of the work of a researcher and its impact on the later literature has been widely studied because of its real and relevant applications, like recruitment, promotion, funding allocation, and so forth. Moreover, with the growing of new on-line digital libraries, like DBLP,[1] Google Scholar,[2] Microsoft Academic Search,[3] CiteSeer[4] etc., which

[1] http://www.informatik.uni-trier.de/ley/db/.

[2] http://scholar.google.com/.

[3] http://academic.research.microsoft.com/.

[4] http://citeseer.ist.psu.edu/.

L. Di Caro (✉) · M. Cataldi · C. Schifanella
Department of Computer Science, University of Torino, Torino, Italy
e-mail: dicaro@di.unito.it

Published online: 26 May 2012

aggregate information about researchers and their outcomes, it is becoming more effective and easier to explore and analyze research products. But, even considering these web services, the pure information about the publication records of an author (like the number of papers and/or citations) often results insufficient for a fair evaluation of scientific profiles because they do not take into account many factors, as the relationships between the authors and their relative scientific influences, which should be directly considered in the evaluation process. In fact, especially when these measures are used for recruitment purposes, it is highly relevant to analyze the scientific dependences among authors in order to estimate the capacity of an author to work and produce research outcomes *without* the people that assisted his/her work until that time. Indeed, in an evaluation process, some co-authored works can unconditionally favour researchers who collaborated with those experts who were able to lead high-quality research projects, while, on the other hand, those who produced such relevant research products could not be distinguished from their co-authors because of the typical assumption of a proportional collaboration among them.

Let us consider for example a case of a young researcher who finishes a Ph.D. program supervised by an expert scientist; considering this situation, we can suppose his/her publication history as a set of few research papers in which the supervisor is involved as co-author. When the researcher applies for some academic or industrial position, the evaluators will be probably interested not only to a pure numerical productivity quantification, which reflects the overall scientific effectiveness of his/her research group, but they will also analyze his/her curriculum with the goal of understanding how much the young scientist would be effective without his/her past research environment. In some sense, based on his/her publication records, the evaluator is interested in analyzing his/her dependence degree on the people who belonged to his/her research group.

The evaluation of such collaboration networks is anyway a challenging task, that, to the best of our knowledge, received few interests so far. Some tools like DBLP Vis[5] or ArnetMiner[6] permit to analyze the academic career of a researcher by visualizing their co-authorship relations, but, nevertheless, they only show the relationships among authors based on the number of co-authored papers and are not suitable for such a challenging task.

Thus, the main goal of this study is to introduce a novel indicator for measuring the dependence among scientists, called *dependence index*, by analyzing their co-authorship network and their shared outputs. This can be obtained through a novel graph-based method that permits to analyze the scientific collaboration between two researchers and estimate their mutual scientific dependences by taking into account the environment in which this collaboration has been growing.

The paper is organized as follows: in "Background" section we first survey the relevant literature on metrics and tools for evaluating collaborations among researchers; then, after describing our overall conjectures and motivations, in "Data and methods" section we introduce the Bipartite Dependence Graph, which models the co-authorship relations that forms the basis for computing the dependence index, *d-index*. Then, after presenting instruments for effective visualization of the resulting information ("Visualization of researchers' collaborations and dependences" section), in "Results" section we report real case scenarios (taken from the web application presented in the "Appendix" section), user studies and analyses of the data that highlight both the validity and the reliability of our research evaluation approach. Conclusions end the paper.

[5] http://dblpvis.uni-trier.de/.

[6] http://arnetminer.org/.

Background

The mechanisms of collaboration among researchers are the focus of a long and established tradition of research (Katz et al. 1997), from qualitative studies on cooperation behaviors (Crane 1969; Chubin 1976; Shapin 1981) to more quantitative approaches (de Beaver and Rosen 1979; Melin and Persson 1996). The latter includes social network-based studies (because of a collaboration network can be easily seen as a social network where people form teams to produce some results), which are generally aiming at understanding the structural determinants and patterns of collaboration (Newman 2001; Barabasi et al. 2002; Moody 2004; Brandao et al. 2007; Wagner and Leydesdorff 2005). Such networks have been deeply analyzed by looking at properties like network topology, size and evolution (Moon et al. 2010; Hou et al. 2008).

The topology of the research network has been also analyzed in Rodriguez and Pepe (2008) and Pepe and Rodriguez (2010), where the authors aim at detecting characteristics like academic department, position, and country of origin from socio-academic networks. In Girvan and Newman (2002), the author presented different analysis of collaboration networks like degrees of separation and giant components.

A slightly different problem concerns the analysis of the individual contributions within a collaboration network (and the individual impact in scientific publication records). In this sense, Hunt (1991), Schmidt (1987), and Verhagen et al. (2003) analyzed this problem by taking into account formal rules about the way scientists collaborate. The individual contribution within a co-authorship network has been also studied in Imperial and Rodríguez-Navarro (2007), where the authors propose a version of h-index that only takes into account the papers in which the author is listed as first (therefore interpretating the authors' order as an explicit statement about the scientific contribution of each researcher).

Using the co-authorship network, several works tried to automatically extract the nature of the existing relationships among authors (with particular attention to the advisor-advisee relations). In Wang et al. (2010), the authors try to discover advisor-advisee relationships from research publication networks by using a machine learning approach from labeled data. Our approach is instead completely unsupervised. In Feeney and Bernal (2010) the authors analyzed the ways in which people seek advice and support from women in their co-authorship networks.

Moreover, in literature many works tried also to analyze the relationship existing among co-authorship and citation networks: in Radicchi et al. (2008) the authors analyze the citation distribution of the papers among different disciplines stating that they follow universal behaviours. In Radicchi et al. (2009) the author aim at discovering the diffusion of scientific credits in the community also relying on citation networks. In Yan and Ding (2009), the authors use four centrality measures (closeness, betweenness, degree and PageRank) within a restricted collaboration network, proving their correlation with the citations count. Chen et al. (2007) applies the PageRank algorithm to collaboration networks as a more effective way to compute scientific influences among works with respect to the citation counts.

Our approach represents, to the best of our knowledge, the first automatic and straightforward technique for mining local dependences in co-authorship networks based on their topologies.

Data and methods

In this section we analyze the problem of evaluating research collaborations proposing a novel metric for estimating the degree of scientific dependence (or dependence degree) among authors based on their bibliographic data records.

Definition 1 (*dependence degree*) considering a researcher a_1 and one of his/her co-authors a_2, we aim to numerically estimate the impact of a_2 within the research output of a_1 by quantifying how much the productivity of all the scientific collaborations of a_1 differ when a_2 is not involved.

This can be evaluated by introducing the concept of productivity, that quantifies, through some existing evaluation metric, the scientific production of a researcher based on his/her co-authored papers.

Definition 2 (*Productivity measure*) We define the productivity of an author (or a set of authors) as a function which quantifies the impact of the set of papers authored by him/her (or by the considered set of authors) on the entire research community (intended as the set of authors who published at least one paper) based on some high-level criteria.

Then, we analyze the authors' relationships taking into account the environment in which they work and produce research outcomes.

Definition 3 (*Co-authorship network*) Considering a researcher (or a set of researchers), we define his/her (or their) co-authorship network (also called in the paper scientific environment or scientific network) as the set of researchers who co-authored at least one paper with him/her (or with all of them, not necessarily within the same work).

Notice that we are not aiming at discovering conceptual dependences (i.e., who is providing the core knowledge to some research output) but we focused our work on numerically quantifying how much the productivity of an author can result dependent on the collaboration with another scientist.

In the next sections, we leverage all these concepts in order to formalize the calculus of the dependence degree.

Quantification of the scientific productivity of authors

In literature there is plenty of methods for evaluating the output of a researcher (either called scientist, or author, in the paper). Most of them consider their publication records as the basis for the scientific evaluation.

In our paper, given an author a_i, we formalize his/her set of research outputs (also called papers, works or outcomes) O_{a_i} as

$$O_{a_i} = \{o_{a_i,1}, o_{a_i,2}, \ldots o_{a_i,n}\}, \tag{1}$$

where $O_{a_i,x}$ is the x-th research output authored, or co-authored, by her. Considering this information, it is possible to quantify the "productivity" of a_i, p_{a_i}, as

$$p_{a_i} = |O_{a_i}|, \tag{2}$$

where $|O_{a_i}|$ is the cardinality of O_{a_i}.

Considering this author productivity evaluation method, it is possible to extend it in order to evaluate the productivity of n authors. In fact given two authors, a_i and a_j, and their research output, O_{a_i,a_j}, defined as

$$O_{a_i,a_j} = O_{a_i} \cap O_{a_j} = \{o_{(a_i,a_j),1}, o_{(a_i,a_j),2}, \ldots, o_{(a_i,a_j),m}\}, \tag{3}$$

where $o_{(a_i,a_j),k}$ is the k-th research output authored by both a_i and a_j, and m is the number of papers authored by both of them, it is now possible to quantify their productivity p_{a_i,a_j} as

$$p_{a_i,a_j} = |O_{a_i,a_j}|. \tag{4}$$

Notice that this approach is extendable to any set of authors with any cardinality. At this point, theoretically, given the entire set of scientific authors, it is possible to model an authors network, *p-Graph*, as an undirected graph that expresses the productivity of each collaboration between two authors as

$$p - Graph = \{V, E, p\}, \tag{5}$$

where $V = \{a_1, a_2, \ldots, a_n\}$ is the complete set of n scientific authors, E is the set of undirected edges, where each $e_{i,j} \in E$ represents an existing collaboration between a_i and a_j (where $a_i, a_j \in V$), motivated by at least one research output co-authored by both of them, and p is the weighting function $(p : E \rightarrow N+)$ representing the *productivity* of a_i and a_j (calculated as explained above).

Considering this formalization, given an author $a_i \in V$, it is possible to formalize the scientific network in which he/she produced research outcomes as

$$Net_{a_i} = \{a_1, a_2, \ldots, a_h\} \tag{6}$$

with $h \leq n$, that is the set of authors directly connected to a_i, in the *p-Graph*, by an edge in E. In other words, $a_k \in Net_{a_i} \Rightarrow \exists\, o_{a_i,a_k}$.

Analyzing scientific dependences between authors through their bipartite dependence graph (*BDG*): the *d*-index

In the previous section we introduced a way to roughly quantify the productivity of n authors by taking into account their research outcomes. But, by itself, p only expresses how fruitful has been the collaboration among the authors in terms of number of co-authored papers, but it is not sufficient for quantifying the scientific dependence among them.

In order to better understand the problem, let us consider a simplified situation as the one shown in Fig. 1. Three authors, *Alice*, *Bob* and *Charlie*, collaborated by publishing several scientific works. In particular, the collaboration between *Alice* and *Bob* (without *Charlie*) resulted in many research outputs as for the collaboration between *Alice* and *Charlie*. On the other hand, the scientific collaboration between *Bob* and *Charlie* did not result in any published work.

This situation can be summarized as follows: two of them (*Bob* and *Charlie*) can be thought as young scientists who are supervised by *Alice*. In this case, *Alice* is leading this small research group and most of the ideas and the necessary expertise can be easily credited to her. Notice that, this fact does not avoid the merits or the contribution of *Bob* and *Charlie* in the considered research outputs. We just state that this situation suggests a sort of scientific dependence of *Bob* and *Charlie* on *Alice*, that is confirmed by the fact that the collaboration between them resulted poorly productive without the collaboration of *Alice* (the only outcome they produced together is, again, co-authored by *Alice*). Considering this example, the scientific dependence of *Bob* and *Charlie* on *Alice* has been highlighted from the analysis of their *co-authorship network* that models the environment (and, therefore, the relationships existing among authors) in which they work. In fact, it permits to analyze the productivity of all pairs of authors with respect to all their co-authors (highlighting, for instance, the fact that *Bob* and *Charlie* were both unable to be so scientifically productive as they were with the support of *Alice*).

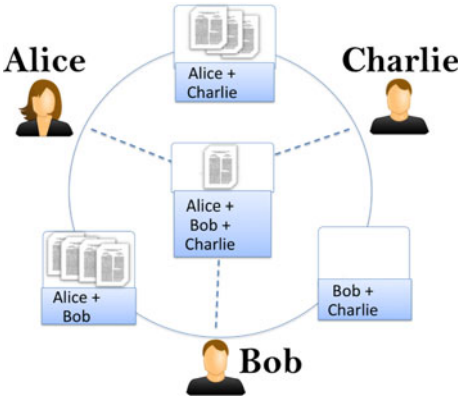


Fig. 1 A simplified example of co-authorship relations: three scientists, *Alice*, *Bob* and *Charlie* published, all together, one paper. However, *Alice* and *Bob* have several publications without *Charlie* and the same happens with *Alice* and *Charlie*. *Bob* and *Charlie* do not have publications without *Alice*

In order to formalize this, we introduce a *Bipartite Dependence Graph*,[7] or *BDG*, that represents two authors in their scientific environment, which is defined by the set of co-authors they have in common. More formally, given two authors a_i and a_j, we define their Bipartite Dependence Graph, BDG_{a_i,a_j}, as

$$BDG_{a_i,a_j} = \{V_{a_i,a_j}, E_{a_i,a_j}, w\} \tag{7}$$

where

- $V_{a_i,a_j} = \{a_i, a_j\} \cup Net_{a_i,a_j}$, where $Net_{a_i,a_j} = Net_{a_i} \cap Net_{a_j}$ is the set of researchers who co-authored at least one paper with both a_i and a_j (not necessarily in the same output);
- E_{a_i,a_j} contains the undirected edges representing collaborations between two authors (i.e., if $e_{a_x,a_y} \in E_{a_i,a_j}$ there is at least one paper co-authored by both a_x and a_y, with $a_x \in \{a_i, a_j\} \wedge a_y \in Net_{a_i,a_j}$);
- w is the weighting function.

Now, we define the weight of each edge within the Bipartite Dependence Graph as the *relative dependence*, w, of a collaboration between two authors with respect to a co-author. In other words, this weight estimates how much the collaboration between a pair of authors is dependent on the collaboration with a common co-author.

In order to better understand this step, let us consider the example shown in Fig. 2. The Bipartite Dependence Graph related to the authors a_1 and a_2 connects two set of verteces: $\{a_1,a_2\}$, which only contains the two considered authors, and Net_{a_1,a_2}, which is the set of their common co-authors. The author a_3, who belongs to the research environment of a_1 and a_2 ($a_3 \in Net_{a_1,a_2}$), represents a researcher who co-authored at least one paper with both a_1 and a_2 (not necessarily in the same work). Thus, the weight of the edge connecting a_1 and a_3 in BDG_{a_1,a_2} represents the dependence of the collaboration between a_1 and a_3 on their collaboration with a_2. This information quantifies how much the collaboration between a_1 and a_3 depends on the scientific contribution of a_2.

More formally, given two authors a_i and a_j and one of their common co-authors a_k, we define the relative dependence of the collaboration between a_i and a_k on a_j as $w^{a_j}_{a_i,a_k}$ in BDG_{a_i,a_j}

[7] A bipartite graph is a graph with two distinct vertex sets such that all edges in the graph are between vertices of different vertex sets.

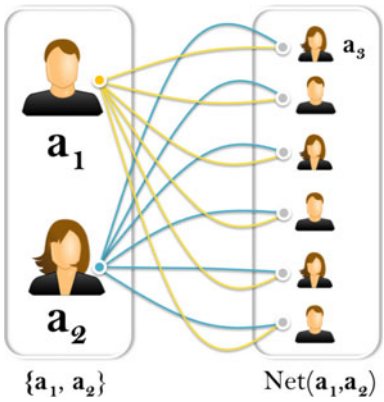


Fig. 2 *Bipartite Dependence Graph*: two co-authors a_1 and a_2 have their shared research network Net_{a_1,a_2} composed by all the co-authors that they have in common

$$w_{a_i,a_k}^{a_j} = 1 - \frac{p_{a_i,a_j,a_k} + p_{a_i,a_k,\neg a_j}}{p_{a_i,a_j,a_k} + p_{a_i,a_k,\neg a_j} + p_{a_j,a_k,\neg a_i}}, \tag{8}$$

where

- p_{a_i,a_j,a_k} is the productivity of the collaboration among a_i, a_j, and a_k, all together;
- $p_{a_i,a_k,\neg a_j}$ is the productivity of the collaboration between a_i and a_k without the contribution of a_j (i.e., the number of research outputs of both a_i and a_k without considering those in which a_j is also involved);
- $p_{a_j,a_k,\neg a_i}$ is the productivity of the collaboration between a_j and a_k without the contribution of a_i (i.e., the number of research outputs of both a_j and a_k in which a_i is not involved).

It is important to notice that the relative dependence $w_{a_i,a_k}^{a_j}$ ranges from *0* to *1*; in particular, $w_{a_i,a_k}^{a_j} \approx 0$ indicates that the dependence of the scientific collaboration between a_i and a_k on a_j is negligible, while a $w_{a_i,a_k}^{a_j} \approx 1$ highlights the contrary. In fact, the second term of the formula increases when the productivity of a_j and a_k, without the contribution of a_i, is negligible ($p_{a_j,a_k,\neg a_i} \approx 0$) and the other cooperations are quantitatively significant ($p_{a_i,a_j,a_k} > 0$ and $p_{a_i,a_k,\neg a_j} > 0$). On the other hand, the higher the $p_{a_j,a_k,\neg a_i}$, the higher the relative dependence.

At this point, given two authors a_i and a_j, and their research network Net_{a_i,a_j}, it is possible to compute the dependence index, *d*-index, of a_i on a_j by analyzing all the relative dependences of the collaborations of a_i (with all the common co-authors) on a_j. In other words, it is now possible to quantify how much the productivity of a_i is dependent on the collaboration with a_j by evaluating all the scientific collaborations of a_i and calculating how much they have been dependent on the scientific contributions of a_j.

More formally, we calculate the dependence index, *d*-index, of a_i on a_j, as

$$d_{a_i}^{a_j} = \frac{p_{a_i,a_j}}{p_{a_i}} \times \frac{\sum_{a_k \in Net_{a_i,a_j}} w_{a_i,a_k}^{a_j}}{|Net_{a_i,a_j}|}, \tag{9}$$

where p_{a_i,a_j} is the productivity of the collaboration between a_i, a_j, and p_{a_i} returns the entire productivity of a_i. It is important to notice that, while the first term permits to numerically quantify the impact of this scientific collaboration over the entire bibliographic record of a_i, the second term calculates the average of the relative dependences of the scientific

collaborations of a_i on a_j. We remember that, since the weights $w^{a_j}_{a_i,a_k}$ range from 0 to 1, also the *d-index* formula lies in the range [0,1] as well. The higher the value of the *d*-index, the higher the scientific dependence of a_i on the collaboration with a_j.

Please notice that $d^{a_j}_{a_i} \neq d^{a_i}_{a_j}$; in fact, even if their scientific network is the same ($Net_{a_i,a_j} = Net_{a_j,a_i}$) their relative dependences with the co-authors can significantly differ, since they are also based on their personal collaboration.

In order to better understand the process, an example (related to the situation summarized in Fig. 1) is shown in Table 1. The common scientific environment of both *Alice* and *Bob* is only formed by *Charlie* and their dependence degrees are therefore estimated based on their collaborations with him. The productivities of the collaborations among both *Alice* and *Bob* with respect to *Charlie* are therefore calculated (Table 1b, c) and provide the base for estimating the overall dependence of *Bob* on *Alice* (the *d*-index = 0.75) and the independence of *Alice* from *Bob* (the *d*-index is 0).

A real, more complex, case is also provided in Table 2. Two authors *A* on *B* collaborated with 5 common co-authors in the way summarized in Table 2a. Based on these collaborations with common co-authors, it is therefore possible to calculate the productivities of each collaboration (Table 2b, c) and estimate their dependence by calculating their *d*-index values.

Normalizing *d*-index values over the entire community: author's dependence coefficient

In "Analyzing scientific dependences between authors through their bipartite dependence graph (*BDG*): the *d*-index" we introduced a novel way to estimate the dependence degree between two authors based on their scientific common network, modeled through a Bipartite Dependence Graph.

These values can now be leveraged to graphically map the scientific dependences of an author as a curve that plots the ordered *d*-index values related to all his/her co-authors. In

Table 1 Summary of the papers co-authored by Alice, Bob and Charlie (a), productivities of the collaborations related to Alice (b) and Bob (c)

(a) Number of papers co-authored

	Alice (A)	Bob (B)	Charlie (C)
Alice (A)	8	5	4
Bob (B)	5	5	0
Charlie (C)	4	0	4

(b) Productivities related to Alice

p_A	$p_{A,B}$	$p_{A,B,C}$	$p_{A,C,\neg B}$	$p_{B,C,\neg A}$	d^B_A
8	5	1	3	0	0

(c) Productivities related to Bob

p_B	$p_{B,A}$	$p_{B,A,C}$	$p_{B,C,\neg A}$	$p_{A,C,\neg B}$	d^A_B
5	5	1	0	3	0.75

The resulting *d*-index values highlight the scientific dependence of Bob on Alice (while the contrary is not true)

Table 2 Real anonymized example; A authored in total 65 papers ($p_A = 65$) while B only 8 ($p_B = 8$)

(a) Number of papers co-authored by A and B w.r.t. their common co-authors.

	Co-author1	Co-author2	Co-author3	Co-author4	Co-author5
A	25	4	6	1	2
B	5	3	1	1	1

(b) Productivities of all the collaborations of A w.r.t. B.

	p_{ABC}	$p_{AC\neg B}$	$p_{BC\neg A}$	w^B_{AC}
Co-author1 (C)	4	21	1	0.0385
Co-author2 (C)	2	2	1	0.2000
Co-author3 (C)	1	5	0	0.0000
Co-author4 (C)	1	0	0	0.0000
Co-author5 (C)	1	1	0	0.0000
d-index (d^B_A)				0.0036

(c) Productivities of all the collaborations of B w.r.t. A.

	p_{BAC}	$p_{BC\neg A}$	$p_{AC\neg B}$	w^A_{BC}
Co-author1 (C)	4	1	21	0.8077
Co-author2 (C)	2	1	2	0.4000
Co-author3 (C)	1	0	5	0.8333
Co-author4 (C)	1	0	0	0.0000
Co-author5 (C)	1	0	1	0.5000
d-index (d^A_B)				0.3176

A and B collaborated by publishing together 5 papers ($p_{AB} = 5$). Their common research environment includes 5 scientists. The productivities of the collaborations within this environment lead to estimate an overall dependence of B on A ($d^A_B = 0.3176$) and the independence of A from B ($d^B_A = 0.0036$)

fact, we believe that such an at-a-glance view, by itself, can give a first sight on this mined knowledge (see also "Dependence curve and comparison among authors" section).

Despite that, we can go further this visual analysis to achieve a one-parameter evaluation system for authors' dependences, as it is the h-index for the number of citations of a researcher. Thus, given the ordered list of d-index values, we calculate the overall author's dependence degree, called *dependence coefficient*, by quantifying how much his/her most important collaborations impact on his/her entire research output. In a sense, we aim to quantify how much a researcher has been able to work without the scientists from whom he/she results more dependent.

More formally, given an author a_i and his/her set of co-authors Net_{a_i}, we calculate the set of his/her co-authors, K_{a_i}, from whom a_i results more scientifically dependent as follows:

1. the system ranks the co-authors in Net_{a_i} in descending order of d-index value;
2. it then computes the *drop* as the first ratio of d-index values between consecutive entries that is higher than the previous one (i.e., it analyzes the dependence curve in order to find the starting point of its long tail);
3. finally, the system selects the co-authors who precede the drop, in the ranked co-authors list, and returns them as K_{a_i}.

At the end of this step, we have associated to the considered author a_i a subset of his/her co-authors $K_{a_i} = \{a_1, a_2, \ldots a_p\}$, where $p \leq |Net_{a_i}|$, from whom he/she is more scientifically dependent.

At this point, we calculate the dependence coefficient of a researcher a_i as his/her dependence degree on the entire set of his/her most influential co-authors K_{a_i}. For this, we apply the method explained in the previous section. In particular, we first define the scientific network of a_i and K_{a_i} as $Net_{a_i,K_{a_i}} = Net_{a_i,a_1} \cup Net_{a_i,a_2} \cup \ldots \cup Net_{a_i,a_p}$, (where $K_{a_i} = \{a_1, a_2, \ldots a_p\}$). Then, for each co-author $a_k \in Net_{a_i,K_{a_i}}$, we calculate the relative dependence of the collaboration between a_i and a_k on the collaboration with the authors in K_{a_i} as

$$w_{a_i,a_k}^{K_{a_i}} = 1 - \frac{p_{a_i,K_{a_i},a_k} + p_{a_i,a_k,\neg K_{a_i}}}{p_{a_i,K_{a_i},a_k} + p_{a_i,a_k,\neg K_{a_i}} + p_{K_{a_i},a_k,\neg a_i}}, \tag{10}$$

where

- p_{a_i,K_{a_i},a_k} is the productivity of the collaboration among a_i, a_k, and at least one author in K_{a_i};
- $p_{a_i,a_k,\neg K_{a_i}}$ is the productivity of the collaboration between a_i and a_k without the contribution of K_{a_i} (i.e., excluding the research outputs of both a_i and a_k in which is also involved at least one author in K_{a_i});
- $p_{K_{a_i},a_k,\neg a_i}$ is the productivity of the collaboration between a least one author in K_{a_i} and a_k without the contribution of a_i (i.e., excluding the research outputs of both a_k and at least one author in K_{a_i} in which is also involved a_i);

At this step, we use all these relative dependences to calculate the dependence coefficient of a_i as

$$c_{a_i} = d_{a_i}^{K_{a_i}} = \frac{p_{a_i,K_{a_i}}}{p_{a_i}} \times \frac{\sum_{a_k \in Net_{a_i,K_{a_i}}} w_{a_i,a_k}^{K_{a_i}}}{|Net_{a_i,K_{a_i}}|}, \tag{11}$$

where $p_{a_i,K_{a_i}}$ is the productivity of a_i in collaboration with K_{a_i} (i.e., the number of research output co-authored by a_i and at least one author in K_{a_i}).

Using the dependence coefficient, we are able to compare researchers focusing on the collaborations that most impacted on their research outputs. In fact, this value permits to estimate how much the productivity of a researcher depends on the scientific collaborators from whom he/she results more dependent. In a sense, through the dependence coefficient, we analyze the impact of the most important scientific collaborations over the entire publication history of the considered researcher.

In the next sections we will show how to use this parameter for evaluation and comparison purposes.

Visualization of researchers' collaborations and dependences

In this section, we present a set of tools that we use within our web application[8] for presenting the information regarding the author's dependences and for allowing the analysis of these data.

[8] presented in "Appendix" section and available at http://d-index.di.unito.it.

Dependence curve and comparison among authors

The first way to analyze the scientific dependences of a researcher is to plot, in ascending order, the *d*-index values related to his/her co-authors within a cartesian coordinate system. In fact, we believe that such an at-a-glance view, by itself, can give a first sight on this mined knowledge. Figure 3 shows an example of what we call *dependence curve*; the curve plots the *d*-index values that form the basis for the calculus of the dependence coefficient. This simple visualization scheme allows the user to give a first insight into the author's dependences, with respect to his/her co-authorship network, by showing in a single chart the ordered *d*-index values with respect to all his/her co-authors.

It is important to notice that this curve can be also positively leveraged as an instrument for comparisons among different authors. In fact the curves permit to compare the scientific dependences of several scientists by focusing on their dependences with respect to their shared collaborators (that represents the common information space in which the comparison is directly possible).

However, considering that the co-authors of two researchers are most of the time different, in order to permit a comparison, we propose an alternative chart where we map onto the *x*-axis a set of meta-authors where each one of them represents one co-author of each considered researcher. More formally, given two authors a_i and a_j and their co-authorship networks $Net_{a_i} = \{a_{i,1}, a_{i,2}...a_{i,n}\}$ and $Net_{a_j} = \{a_{j,1}, a_{j,2}...a_{j,m}\}$ (where $a_{j,t}$ is the *t*-th co-author of a_j ordered by the *d*-index values, and *n* is not necessarily equal to *m*), we map onto the *x*-axis the set of meta-authors $Net_{meta} = \{a_{k,1}, a_{k,2}...a_{k,r}\}$ where each author $a_{k,h} \in Net_{meta}$ represents one co-author of each considered researcher (i.e., $a_{k,h}$ represents $a_{i,n-r+h}$ and $a_{j,m-r+h}$ respectively), and $r = m$ if $m \leq n$, otherwise $r = n$.

In this way, we basically compare scientists based on their highest dependence values instead of their shared co-authors (that, obviously, not necessarily exist). In a sense, we compare researchers based on their absolute dependence values without taking into account from whom they are dependent. An example is shown in Fig. 4.

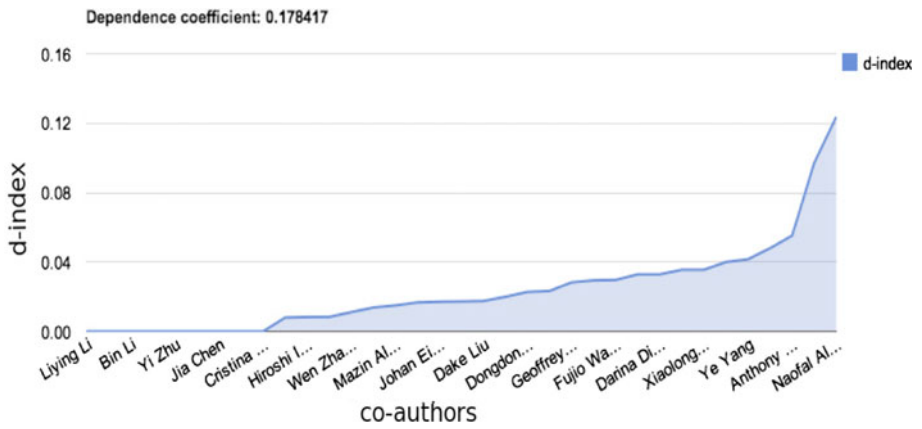


Fig. 3 The dependence curve: given an author, the *d*-index values related of his/her co-authors are shown through a simple chart (where the *x*-axis represents the co-authors, ordered based on their *d*-index value, and the *y*-axis reports the related *d*-index). The curve therefore represents the ordered list of *d*-index values (the dependence coefficient is reported above the curve)

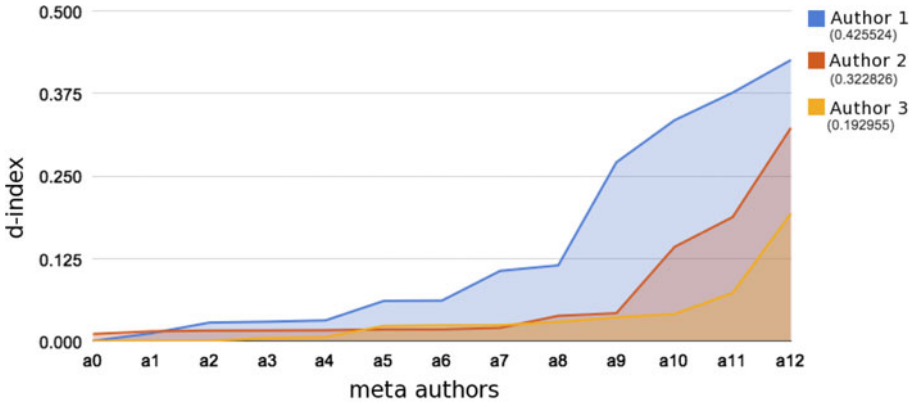


Fig. 4 Comparison without shared co-authors: onto the x-axis are mapped meta-author's labels and each one of them represents one co-author of each considered researcher

Please notice that, in order to clarify when a comparison is statistically significant, the web application also reports the analysis of the variance on the related dependence patterns, performed through the ANOVA test (Fisher 1918).

Mutual Dependence Curve

While the previous visualization scheme provides a simple way to analyze the dependences of a researcher, it does not permit to evaluate the reciprocal ones. In fact, given two scientists, it could be difficult to establish who of them is more dependent on the other by only analyzing their curves. For this reason, we introduce the *Mutual Dependence Curve* that, given an author a_i, plots all the co-authors onto the x-axis and, for each co-author a_k, shows their *mutual dependence* calculated as $d_{a_i}^{a_k} - d_{a_k}^{a_i}$ onto the y-axis. The co-authors are ordered based on these values.

Using this simple visualization method it is possible to infer how much the collaboration is overbalanced and who of them is more dependent on the other. In fact, if $d_i^k - d_k^i > 0$ the author a_i is more dependent on a_k than a_k on a_i, and vice versa. Thus, with the Mutual Dependence Curve, the user can view the collaborators who depend on the considered researcher (those plotted under the x-axis) and the authors that more likely lead the scientific collaborations with the considered researcher (those plotted above the x-axis). An example is shown in Fig. 5.

Collaboration Map

Since researchers tend to work in communities, we provide a tool, called *Collaboration Map*, for analyzing scientific dependences among authors with respect to the communities they form. Thus, in this chart, we aim to plot the co-authors based on the *local communities* they form with respect to the considered researcher. For this, instead of equally distribute the co-authors onto the x-axis, we place them according to their productivity with respect to the considered author. The idea is to plot closer, onto the x-axis, the co-authors that collaborated more prolifically with him/her within his/her research environment.

In order to do that, given an author a_i, we compute his/her square collaboration matrix C_i where both rows and columns represent the co-authors of a_i, and $C_i[j][k] = p_{a_i,a_j,a_k}$ is the

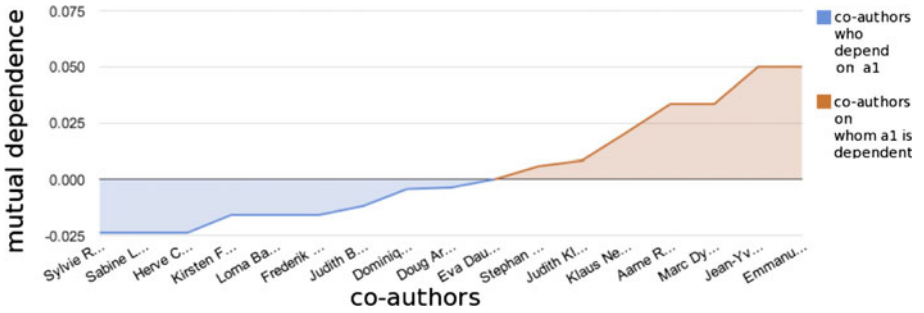


Fig. 5 *Mutual Dependence Curve*: the x-coordinate represents the co-authors of the considered researcher. The y-coordinate shows the reciprocal dependences among the researcher under evaluation and his/her co-authors. The collaborators who depend on the considered author are plotted under the x-axis (and visualized in *blue*) while the authors on who he/she depends are plotted above the x-axis (and visualized in *red*)

productivity of a_i with a_j and a_k (the diagonal $C_i[j][j] = p_{a_i,a_j}$ simply reports the productivity of a_i with a_j). From this, we compute a dissimilarity matrix of co-authors of a_i, D_i, where

$$D_i[j][k] = 1 - cos(C_i[j], C_i[k]). \tag{12}$$

where the function *cos* measures the cosine similarity between two vectors, $C_i[j]$ and $C_i[k]$, representing, respectively, the j-th and k-th rows of the Collaboration Matrix of the author a_i.

We then apply the clustering approach presented in Pelleg and Moore (2000), the X-Means (which is able to automatically determine the optimal number of clusters), to the dissimilarity matrix in order to find clusters of co-authors who collaborated prolifically among them with respect to considered author.

At this point, in order to visualize this information we need to map it into a one-dimensional ordering (the x-axis). For this, without loss of generality, we use multi-dimensional scaling (MDS) proposed in Torgerson (1958) to embed the co-authors onto a 1-dimensional order. MDS works as follows: given as inputs (1) a set of N objects, (2) a matrix of size $N \times N$ containing pairwise distance values and (3) the desired dimensionality k, MDS tries to map each object into a point in the k-dimensional space in such a way to minimize the *stress* value, defined as

$$stress = \sqrt{\frac{\sum_{i,j}(dist'_{i,j} - dist_{i,j})^2}{\sum_{i,j}(dist_{i,j})^2}}, \tag{13}$$

where $dist_{i,j}$ is the actual distance between two objects obj_i and obj_j and $dist'_{i,j}$ is the distance between the corresponding points in the resulting k-dimensional space. Therefore, by providing as input $|Net_{a_i}|$ co-authors, their distances defined through the dissimilarity matrix D, and $k = 1$ target dimension, the resulting order of co-authors would preserve the distances (i.e., the dissimilarities) among co-authors as best as possible. The resulting MDS value is used for mapping each co-author onto the x-axis. Then, using the retrieved set of clusters, we display the co-authors by differently coloring them based on the clusters they form.

Moreover, in order to help the user distinguish the co-authors based on their dependences on the considered researcher, we map them onto y-axis (as done for the Mutual Dependence Curve) by using their mutual dependences on the considered researcher.

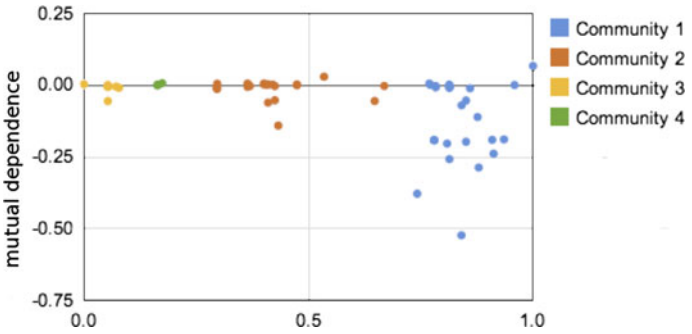


Fig. 6 *Collaboration Map*: this tools permits to discover latent research communities, displayed with different *colors*, among the co-authors of a considered research. The *x*-coordinate represents the first dimension of the Multi-Dimensional Scaling applied on the dissimilarity matrix among the co-authors and the *y*-coordinate shows the reciprocal dependences among the researcher under evaluation and his/her co-authors. In this example, the author seems as one of the leaders of all the four communities (because most of the co-authors are more dependent on his/her than the contrary)

Thus, using the Collaboration Map, it is possible to graphically identify the local communities of the considered researcher (i.e., closer the co-authors onto the *x*-axis, more prolific their collaboration within the co-authorship network), and to analyze the mutual scientific dependence of each co-author on the considered researcher (i.e., the lower his/her *y*-value, the more the scientific dependent on the considered researcher). Thus, the collaborators who depend on the considered researcher are plotted under the x-axis, while the authors on whom the considered researcher depends are visualized above the x-axis. An example is shown in Fig. 6.

Results

Differently from the common approach of evaluating researchers according to some quality/quantity measure regarding their scientific outputs, in this paper we have shown how to focus the evaluation on the analysis of their scientific collaborations. In this section we provide real, high-level, scenarios, user studies and analyses that can help understand the introduced evaluation schema and motivate the benefits of using the proposed approach.

For our experiments, we considered the whole DBLP bibliographic data set[9] containing 1,097,418 scientific authors and 1,924,030 papers.[10] All the data are accessible through the web application available at http://d-index.di.unito.it, presented in "Appendix" section.

Case study

In this section, in order to analyze the proposed evaluation metric, we analyze the real cases of well-known scientists (as in Sidiropoulos et al. (2007)) within the scientific community:

[9] http://www.informatik.uni-trier.de/ley/db/.

[10] The data set is updated on March 2012. Please notice that, within our system, we rely on the disambiguated authors name provided by Tang et al. (2007). This name disambiguation system is based on a constraints probabilistic model and aims at finding, extracting, and fusing the semantics-based profiling information of a researcher from the Web.

Table 3 The list of co-authors, ordered by their *d*-index values related to Dr. Faloutsos (a) and Dr. Leskovec (b), and the number of papers they co-authored with them

(a)

Position	Faloutsos' co-authors	*d*-index	# Co-authored papers
1st	Caetano Traina Jr.	0.0241	22
2nd	Agma J. M. Traina	0.0196	23
3rd	Timos K. Sellis	0.0166	10
4th	Jimeng Sun	0.0127	23
…	…	…	…
23th	Jure Leskovec	0.0052	18

(b)

Position	Leskovec's co-authors	*d*-index	# Co-authored papers
1st	Christos Faloutsos	0.0785	18
2nd	Jon M. Kleinberg	0.0657	13
3rd	Michael W. Mahoney	0.0286	5
4th	Daniel P. Huttenlocher	0.0286	7
…	…	…	…

- Dr. Christos Faloutsos: he is Professor of Computer Science at Carnegie Mellon University and his research interests include data mining for streams and networks, indexing for multimedia and bio-informatics databases, and performance.[11]
- Dr. Jure Leskovec: he is one of the most emerging scientists in the data mining community. He is currently Assistant Professor of Computer Science at Stanford University.[12] He completed his Ph.D. program under the supervision of Dr. Christos Faloutsos in September 2008.

Both scientists produced many research outcomes; in detail, Dr. Faloutsos published more than 300 papers, by also collaborating with 345 different researchers, while Dr. Leskovec produced 72 papers with 72 different co-authors. Together, they published 18 papers in different venues.

Using the dependence index, it is now possible to calculate their dependences on their co-authors (Table 3a, b) and their related dependence curves, that can be positively leveraged to compare their dependence patterns (Fig. 7). As it was easily presumable from their publication records, the curves highlight a higher scientifically independence of Dr. Faloutsos with respect to Dr. Leskovec.

Moreover, through the analysis of the *d*-index values shown in Table 3a, b, Dr. Leskovec resulted $\sim$10 times more dependent on the collaboration with Dr. Faloutsos than the contrary (0.0785 vs 0.0052).

It is important to notice that, even if this result could appear as highly predictable from the pure productivities (Dr. Faloutsos represents for Dr. Leskovec the most productive scientific collaboration, while the contrary is not true), the number of co-authored papers is not directly proportional to the *d*-index value. In other words, a higher number of

[11] Information updated on March 2012.

[12] Information updated on March 2012.

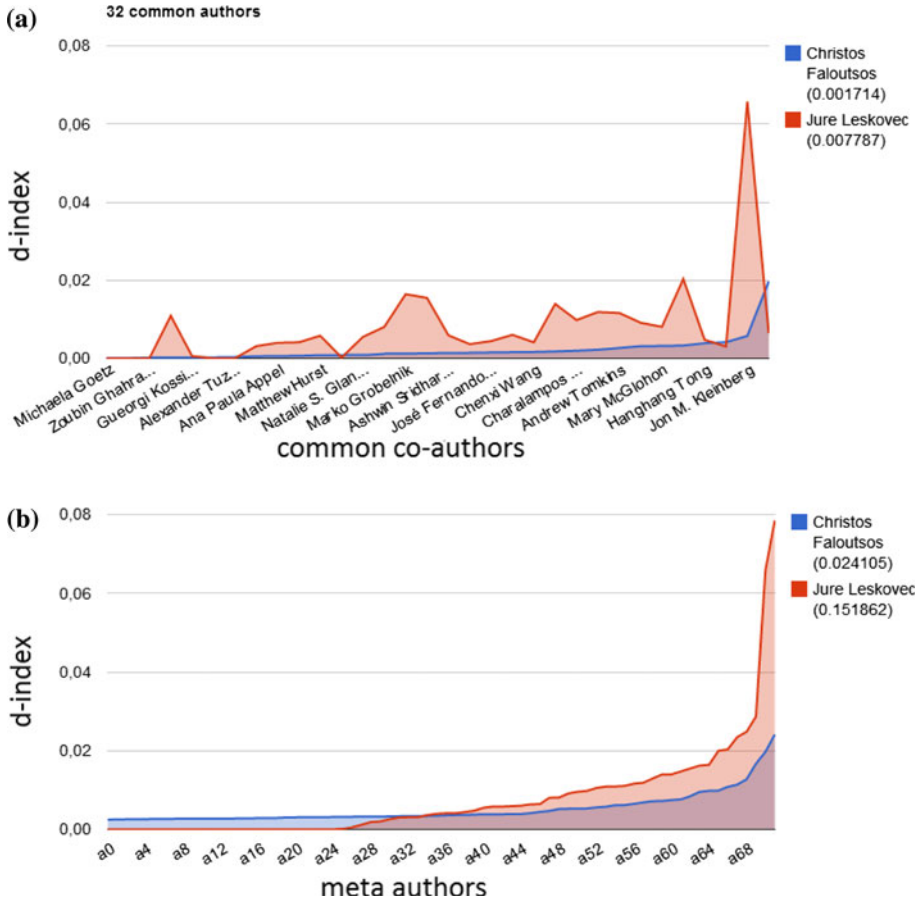


Fig. 7 The comparison between the dependence curve of Dr. Christos Faloutsos and the curve of one of his ex Ph.D. student, Dr. Jure Leskovec, based on their shared co-authors (**a**) and their highest dependence values (**b**)

co-authored papers does not necessarily produce a higher *d*-index value. In fact, as it is possible to notice from Table 3a, Dr. Faloutsos resulted, for example, more dependent on Dr. Timos K. Sellis than from Dr. Leskovec, despite the numbers of co-authored papers are lower (10 vs 18). This is obtained because, even if the collaboration between Dr. Faloutsos and Dr. Leskovec resulted highly fruitful, there were many other papers published by him, without the collaboration of Dr. Leskovec, co-authored by some of their common co-authors. On the other hand, the scientific roles of Dr. Timos K. Sellis resulted more crucial for Dr. Faloutsos because his "incapacity" to maintain the same productivity, without his collaboration, within their common research environments.

In other words, the *d*-index analyzes the scientific collaboration network to retrieve those authors that resulted more crucial for the maintenance of fruitful collaborations among researchers. Indeed, the most general assumption of our method is that, if an author a_i, after a very productive collaboration with a_j, was not able to maintain a similar productivity with their common co-authors, it suggests a sort of dependence on a_j that is motivated by this relaxation of the scientific collaborations with the scientists that usually worked with him/her.

Notice that also the dependence coefficient is not strictly dependent on the total number of published works. In fact, even if the trends of these values highlight an overall correlation (see also "Mutual dependence curve" section), a higher number of research outcomes does not automatically imply a lower dependence coefficient. For example, let us consider the case of Dr. Philip S. Yu, Professor at the University of Illinois, who is the most productive author within the considered data set (667 works). In Fig. 8 we compare his dependences with the ones of Dr. Faloutsos and Dr. Leskovec: the dependence curve of Dr. Yu resulted similar to the one of Dr. Faloutsos, who "only" authored 346 papers. Thus, even if the number of considered papers was significantly different (667 vs 346), the comparison between their curves established a sort of draw, highlighting how this metric is not strictly conditioned by the total number of considered research outcomes.

Analyses of the overall trends about the entire community

In the previous section we reported real case studies that help understand the meaning of the introduced *d*-index metric and the use of the visualization tools provided by our web environment.

Now we leverage all the data regarding the entire scientific community to study its overall trend with respect to the introduced *d*-index measure. In particular we analyze the relation between the dependence coefficient and three different parameters: the total number of published papers (Fig. 9a), the total number of co-authors (Fig. 9b) and the length of the scientific career (calculated as the period, expressed in years, from the first publication of an author to his/her last one, Fig. 9c).

In Fig. 9a each researcher within the considered dataset of 1,097,418 scientific authors has been mapped by using his/her total number of papers and his/her dependence coefficient. As it is clearly understandable, the highest dependence coefficients have been obtained by those authors who only co-authored few papers, while those scientists who were able to produce a significant number of outcomes have been associated with a very low dependence coefficient. Notwithstanding, as already mentioned in "Case study", the dependence coefficient of an author is not strictly correlated to his/her total number of papers; for example, a low number of published works does not necessarily mean a very

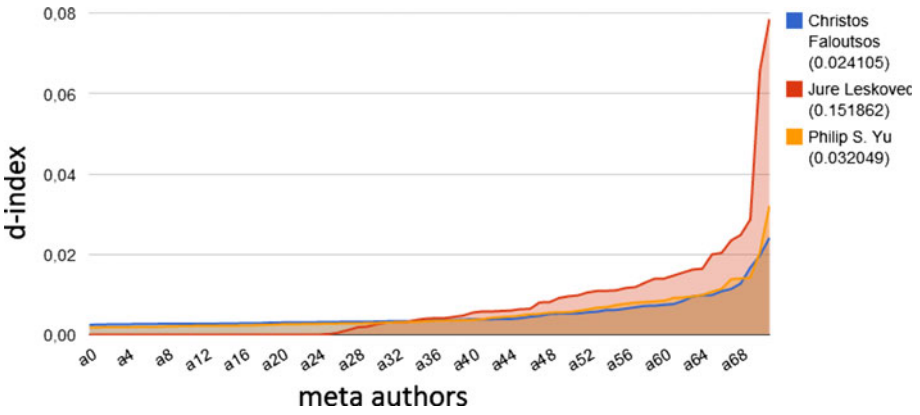


Fig. 8 The comparison among the dependence curve of Dr. Christos Faloutsos (*blue curve*), Dr. Jure Leskovec (*red curve*) and Dr. Philip S. Yu (*in yellow*). Dr. Philip S. Yu is the author, within the considered data set, who co-authored the maximum number of works (667 papers)

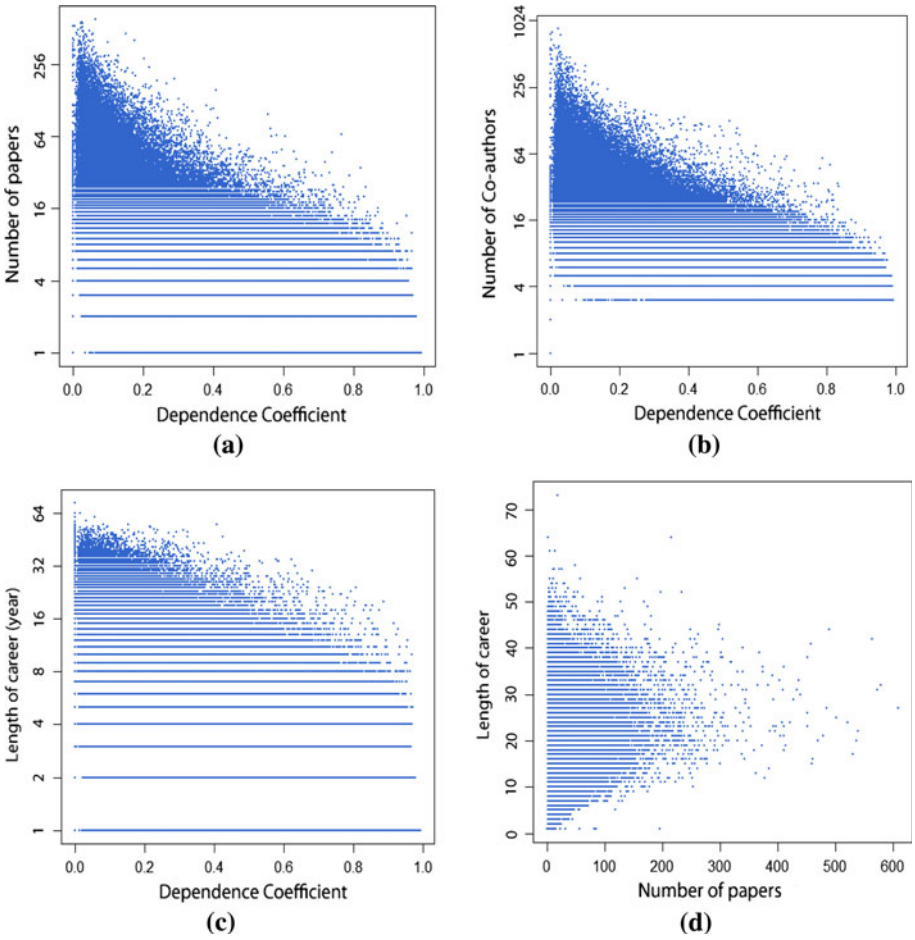


Fig. 9 The relation, over the entire DBLP community, between the dependence coefficient and the total number of published papers (**a**), the total number of co-authors (**b**) and the length of the scientific career (**c**). These parameters have been mapped onto the *y*-axis by using a logarithmic scale. Finally, in **d**, we show the relation over the entire DBLP community between the number of published papers and the length of the career

high dependence coefficient (in the Fig. 9a, the dependence coefficients of the authors with a number of papers lower than 5 range from 0 to 1). In fact, within our system, a young researcher who co-authored very few papers with scientists belonging to different communities (i.e., his/her co-authors never or rarely collaborate among themselves), already proved a sort of independence that is demonstrated by his/her capacity to work, maintaing the same productivity, with people belonging to different research groups.

Therefore, even if the trend of the dependence coefficients seems to be inverse to the one of the total number of published papers (the higher the number of research outcomes, the lower the dependence coefficient), these values are not strictly inversely proportional.

Similar considerations are possible regarding the number of co-authors; in Fig. 9b each researcher is represented by using his/her total number of co-authors (i.e. the number of authors with whom he/she produced at least one research outcome) and his/her dependence

Table 4 The number of users that replied to our questionnaire, grouped based on their academic position

Academic position	Avg. # of papers	Cardinality
Ph.D. students	7.5	6
Short-term researchers	10.45	11
Permanent researchers/assistant professors	22.6	16
Professors	52.0	15

coefficient. Again, a lower number of co-authors does not necessarily lead to a higher dependence coefficient.

Finally, in Fig. 9c we plot the dependence coefficients with respect to the length of the scientific career of the authors (calculated as the arithmetical difference of the year of the last publication minus the year of the first one). As it is possible to notice, these values seem again inversely proportional (the longer the career, the lower the dependence coefficient).

It is interesting to notice that, within this data set, this result could lead to believe in a direct correlation between the length of the career and the total number of published papers (and, therefore, a lower dependence coefficient value). However as shown in Fig. 9d, these two parameters seem not correlated.[13] In fact, the overall independence of an older scientist is not given from his/her pure number of research outcomes, but it is correlated with his/her relationships with the co-authors. Indeed, a scientist who prolonged his/her career for many years, necessarily changed many research collaborators. Therefore, his/her low dependence coefficient is due to the fact he/she was able to change collaborators and keep publishing without suffering a drop in his/her publication record.

User study

In order to evaluate how the *d*-index measure is able to capture dependences among scientific collaborators, we conducted a user study by asking 87 scientists, with different backgrounds and education and belonging to the different departments from different countries, to answer a questionnaire about their scientific dependences. Among all, 48 scientists replied to our call (i.e., about the 55%). A detailed view of the users set is shown in Table 4.

The users answered a questionnaire in which they were asked to order three randomly-picked co-authors according to their feeling of "scientific dependence" on them. We repeated this phase four times for each user (therefore obtaining, for each of them, four rankings of four different subset of co-authors).

Then, in order to asses the validity of our approach in measuring the dependences among scientific collaborators, we compared the users rankings with the ones provided by our *d*-index. This comparison has been performed by using the Kendall's Tau coefficient (Kendall 1938) and the Goodman and Kruskal's Gamma (Goodman and Kruskal 1954).

[13] As it results from Fig. 9d, within our data set, a scientist with a long career (for example of $\sim$50 years) owns, in average, a lower number of co-authored works with respect to many younger scientists. Analyzing the reason of this result is out of the scope of this paper, but some hypotheses are possible. For example, this could lead to believe in a significant improvement in the rhythm of the publication process along the last decades. On the other hand, an alternative explanation can be that the coverage of DBLP significantly improved in the recent years.

Table 5 The results of the user study in terms of Kendall's Tau coefficient and Goodman and Kruskal's Gamma. The users have been grouped based on their academic positions

Academic position	Kendall's Tau	G. & Kruskal's Gamma
Ph.D. students	0.833	0.851
Short-term researchers	0.686	0.696
Permanent researchers/assistant professors	0.662	0.679
Professors	0.722	0.777
Overall	0.691	0.712

Using these measures it is possible to capture the rank correlation between different orderings. In fact, the Kendall's Tau coefficient calculates the distance between the two orderings by relying on the number of pairs that are concordant and discordant in the considered rankings without taking into account the ties (items that actually have the same *d*-index value). Since in our case we may have a significant number of ties, we also calculate the Goodman and Kruskal's Gamma, which permits to test the strength of the association of two different rankings when both are measured at the ordinal level (by making no adjustment for either data size or ties). This measure is known to be more appropiate than Kendall's Tau when the data contain many tied observations.

The results are calculated as the average of the Kendall's Tau coefficients (and the Goodman and Kruskal's Gamma) on the rankings provided by each user, and they are shown in Table 5. As it is possible to notice, this experiment highlights both the validity of our approach (an overall average Kendall's Tau coefficient of 0.691 and Goodman and Kruskal's Gamma of 0.712, which prove a positive correlation among the orderings provided by the users and the ones elaborated through the *d*-index measure) and its reliability; in fact, the resulting Kendall's Tau coefficients are similar for all the considered academic categories (0.833 for the Ph.D. Students, 0.686 for the short-term researcher, 0.662 for permanent researcher and/or Assistant Professor, and 0.722 for Professors), therefore resulting independent by the considered category of users. Note that the values calculated with the Goodman and Kruskal's Gamma show an improvement with respect to the former.

Finally, in order to better test the reliability of the proposed dependence evaluation index, we also grouped the considered users based on different parameters: in particular, we analyzed the average Kendall's Tau coefficients and Goodman and Kruskal's Gamma values for the groups of users having similar behaviour with respect to the total number of published papers and the total number of co-authors. The results are shown in Tables 6 and 7 and they are consistent with the ones provided in Table 5 (the average coefficients and gammas are always higher than 0.6).

In conclusion, considering the presented user study, our proposed measure resulted both valid and reliable, demonstrating to be effective, from a user's point of view, for all the presented categories of users.

Conclusions and future works

The problem of evaluating the quality of researchers' outputs has been broadly studied whereas few works attempted to discover local dependences among researchers in case of co-authored papers. In this sense, we proposed a novel measure, the *d-index*, which aims at

Table 6 The results of the user study in terms of Kendall's Tau coefficient and Goodman and Kruskal's Gamma

Number of users	Number of papers	Kendall's Tau	G. & Kruskal's Gamma
22	From 3 to 20	0.712	0.727
19	From 20 to 50	0.672	0.672
7	From 50 to 120	0.698	0.777

Within this experiment, the considered users have been grouped based on their total number of co-authored papers

Table 7 The results of the user study in terms of Kendall's Tau coefficient and Goodman and Kruskal's Gamma

Number of users	Number of co-authors	Kendall's Tau	G. & Kruskal's Gamma
29	From 5 to 20	0.689	0.733
15	From 20 to 50	0.696	0.701
4	From 50 to 100	0.722	0.722

Within this experiment, the considered users have been grouped based on their total number of collaboration within their entire research careers

uncovering dependence degrees among the authors according to their research environment and their publication history. We then presented several visualization tools for analyzing and comparing scientific collaborations of all the researchers in the DBLP bibliographic database. Finally, we provided case and user studies that help understand our evaluation scheme.

However, some limitations emerged from the study of this challenging task; in particular, the visualization of the dependences of a researcher is complex (especially when each author maintains hundreds of scientific collaborations) and the dependence curve can result uneffective to graphically highlight and summarize dozens of collaborations. Moreover, the comparison among researchers based on their curves can result hardly understandable, because most of the time they simply do not share the same data (i.e., two researcher do not share the same collaborators).

Considering these problems, our future research will include a more accurate testing process of the proposed index (by extending the user study in order to analyze other aspects of the introduced d-index measure) and alternative visualization tools to help the user analyze dependences among researchers. Moreover, we plan to extend our work in order to analyze collaborations among entire research groups and evaluate their relative dependences based on their co-authored works. In particular, we aim to study how these collaborations evolves and how much they are dependent on the collaboration among individual researchers (for example, the leaders of the groups). We also plan to extend the work in order to include alternative productivity measures (the number of citations, or the h-index of the shared research outputs) and evaluate the impact of existing social rules that can condition their quantifications. In conclusion, in order to improve the usability of the proposed web application, we plan to carefully analyze the disambiguation problem of the authors' names, and provide alternative methods to help the user solve the homonymy problem.

Fig. 10 Different visualizations of the career of Dr. Christos Faloutsos taken from http://d-index.di.unito.it; these tools permits to visualize the information about the author (**a**), his dependence curve (**b**), his mutual dependence curve (**c**), a comparison with another scientist (**d**) and his Collaboration Map (**e**), the relations between his dependence coefficient and his total number of papers (**f**), number of his co-authors (**g**), and length of his career (**h**) compared with the whole scientific community

Appendix: The web application: http://d-index.di.unito.it

In this section we present a web environment[14] for evaluation of dependences among scientific researchers. Within the presented web application, it is possible to perform the following operations:

- visualization of scientific profiles;
- graphical analyses of the scientific dependences of the authors on all their co-authors;
- graphical analyses of the mutual dependences of the each researcher relatively to his/her co-authors;
- comparison among authors, based on their dependences on their co-authors (shared or not);
- analysis of the dependences of an author by taking into account the local research communities in which he/she has been involved (intended as groups of co-authors who, together, have frequently collaborated with him/her).
- comparisons with the entire community, focusing on the dependence coefficient, with respect to the total number of papers, total number of co-authors and length of career.

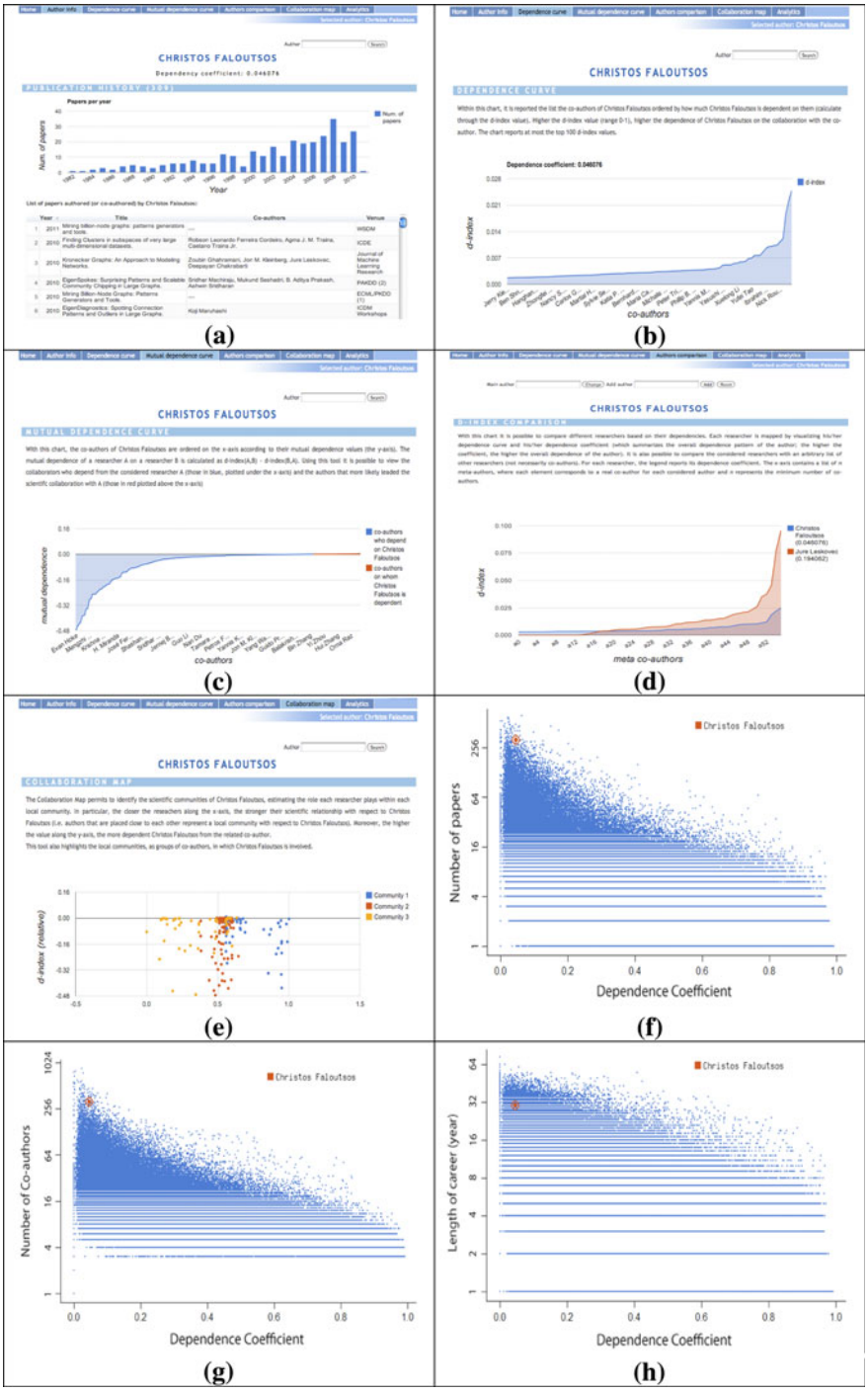
Figure 10 shows a set of screenshots taken from http://d-index.di.unito.it. Within this system a user can search for authors and analyze their scientific profiles, according to the proposed measures. In particular, given a researcher's name, the system permits to retrieve the related authors (by querying the database with the inserted search string) and to disambiguate the resulted authors' list.[15] The user can therefore analyze the scientific profile of the selected author (Fig. 10a) by studying his/her scientific career (visualized as a histogram of papers per year), the complete list of scientific outcomes (through the standard information as the title, the name of the conference/journal, the year of the publication, etc.) and his/her co-authors over the entire career.

Then, as shown in Fig. 10b, it is possible to analyze the 2-dimensional dependence curve of the selected scientist that visualizes all his/her scientific dependences (plotted as explained in "Dependence curve and comparison among authors"). As already introduced, the curve permits to graphically order his/her co-authors with respect to their *d*-index value; in other words, this chart allows the user to easily visualize the authors from whom the considered author is more scientifically dependent. In addition, the user can analyze his/her scientific relationships with his/her co-authors by visualizing his/her Mutual Dependences Curve (Fig. 10c), graphically separating the authors who depend on the considered researcher (in blue) from those he/she is dependent on (in red).

Using the dependence curve, it is also possible to compare different authors in terms of their dependences (Fig. 10d) based on shared co-authors (in case they exist) or their highest dependence values (mapped through meta authors). The user can compare, within the same chart, an unlimited number of authors.

[14] available at http://d-index.di.unito.it.

[15] Please notice that, for the authors' name disambiguation process, we leverage the authors' list provided by Tang et al. (2010); thus, the system is not currently able to distinguish among scientists who are not disambiguated within the considered data set.


CHRISTOS FALOUTSOS
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
Number of papers
Dependence Coefficient
Number of Co-authors
Length of career (year)
Christos Faloutsos

Finally, as shown in Fig. 10e, it is possible to get an alternative view of the scientific dependences of the author by taking into account the communities (intended as group of co-authors) in which he/she is involved. This is possible through the use of the Community Map (explained in "Collaboration map" section). Using this chart, it is possible to organize the co-authors not only with respect to the dependence on them, but also based on the local communities they form with respect to their collaboration with the considered author. In this way it is possible to estimate the number of research groups in which the author is involved (visualized with different colors) and, for each of them, the relative dependence of each researcher belonging to it on the considered author.

Finally, the application permits to compare the researcher with the entire community by focusing on the dependence coefficient with respect to different parameters (total number of outcomes, total number of co-authors and length of the scientific career). Within these charts, all the researchers within DBLP are mapped with blue points, whereas the considered researcher is highlighted in red. Moreover, within these charts, it is possible to map multiple authors and compare them with respect to the entire community. Some examples are shown in Fig. 10f, g, h.

References

Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications, 311*(3–4), 590–614.

Brandao, W. C., de Oliveira e Silva, A. B., & Parreiras, F. S. (2007). Social networks in information science: Behavioral evidences of the researchers and evolutional trends of the coauthor network. *Inf Inf 12*(Special Issue), Portuguese.

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with google's pagerank algorithm. *Journal of Informetrics, 1*(1), 8–15.

Chubin, D. E. (1976). The conceptualization of scientific specialties. *The Sociological Quarterly, 17*(4), 448–476.

Crane, D. (1969). Social structure in a group of scientists: A test of the "invisible college" hypothesis. *American Sociological Review, 3*, 335–352.

de Beaver, D., & Rosen, R. (1979). Studies in scientific collaboration. *Scientometrics, 1*(2), 133–149.

Feeney, M., & Bernal, M. (2010). Women in stem networks: Who seeks advice and support from women scientists?. *Scientometrics, 85*(3), 767–790.

Fisher, R. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh, 52*(2), 399–433.

Girvan, M., & Newman, M. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America, 99*(12), 7821.

Goodman, L., & Kruskal, W. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732–764.

Hou, H., Kretschmer, H., & Liu, Z. (2008). The structure of scientific collaboration networks in scientometrics. *Scientometrics, 75*(2), 189–202.

Hunt, R. (1991). Trying an authorship index. *Nature, 352*(6332), 187.

Imperial, J., & Rodríguez-Navarro, A. (2007). Usefulness of Hirschs h-index to evaluate scientific research in Spain. *Scientometrics, 71*(2), 271–282.

Katz, J. S., Katz, J. S., Martin, B.R., & Martin, B. R. (1997). What is research collaboration?. *Research Policy, 26*, 1–18.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika, 30*(1/2), 81–93.

Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics, 36*, 363–377.

Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review, 69*(2), 213–238.

Moon, S., You, J., Kwak, H., Kim, D., & Jeong, H. (2010). Understanding topological mesoscale features in community mining. In *Communication systems and networks (COMSNETS), 2010 second international conference on* (pp. 1–10). Bangalore, India: IEEE.

Newman, M. E. J. (2001). Scientific collaboration networks: I. network construction and fundamental results. *Physical Review E 64*, 016131.

Pelleg, D., & Moore, A. W. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the seventeenth international conference on machine learning* (pp. 727–734). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Pepe, A., & Rodriguez, M. (2010). An in-depth longitudinal analysis of mixing patterns in a small scientific collaboration network. *Scientometrics, 85*(3).

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences, 105*(45), 17268.

Radicchi, F., Markines, B., & Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E, 80*(5), 056103.

Rodriguez, M., & Pepe, A. (2008). On the relationship between the structural and socioacademic communities of a coauthorship network. *Journal of Informetrics, 2*(3), 195–201.

Schmidt, R. (1987). A worksheet for authorship of scientific articles. *Bulletin of the Ecological Society of America, 68*(1), 8–10.

Shapin, S. (1981). Laboratory life. the social construction of scientific facts. *Medical History, 25*(3), 341–342.

Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics, 72*(2), 253–280.

Tang, J., Zhang, D., & Yao, L. (2007). Social network extraction of academic researchers. In *Proceedings of the 2007 seventh IEEE international conference on data mining* (pp. 292–301). Washington, DC, USA: IEEE Computer Society.

Tang, J., Yao, L., Zhang, D., & Zhang, J. (2010). A combination approach to web user profiling. *ACM Transactions on Knowledge Discovery from Data (TKDD), 5*(1), 2.

Torgerson, W.S. (1958). *Theory and methods of scaling*. Malabar, FL: R.E. Krieger Pub. Co.

Verhagen, J., Wallace, K., Collins, S., & Scott, T. (2003). QUAD system offers fair shares to all authors. *Nature, 426*(6967), 602.

Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy, 34*, 1608–1618.

Wang, C., Han, J., Jia, Y., Tang, J., Zhang, D., Yu, Y., & Guo, J. (2010). Mining advisor-advisee relationships from research publication networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 203–212). San Diego, CA, USA: ACM.

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology, 60*(10), 2107–2118.