D-INDEX: a web environment for analyzing dependences among scientific collaborators

Claudio Schifanella, Luigi Di Caro, Mario Cataldi, Marie-Aude Aufaure

D-INDEX: a Web Environment for Analyzing Dependences among Scientific Collaborators

Claudio Schifanella
Università di Torino
Torino, Italy
schi@di.unito.it

Luigi Di Caro
Università di Torino
Torino, Italy
dicaro@di.unito.it

Mario Cataldi
École Centrale Paris
Paris, France
mario.cataldi@ecp.fr

Marie-Aude Aufaure
École Centrale Paris
Paris, France
marie-aude.aufaure@ecp.fr

ABSTRACT

In this work, we demonstrate a web application, available at *http://d-index.di.unito.it*, that permits to analyze the scientific profiles of all the researchers indexed by DBLP[1] by focusing on the collaborations that contributed to define their curricula. The presented application allows the user to analyze the profile of a researcher, her dependence degrees on all the co-authors (along her entire scientific publication history) and to make comparisons among them in terms of dependence patterns. In particular, it is possible to estimate and visualize how much a researcher has benefited from collaboration with another researcher as well as the communities in which she has been involved. Moreover, the application permits to compare, in a single chart, each researcher with all the scientists indexed in DBLP by focusing on their dependences with respect to many other parameters like the total number of papers, the number of collaborations and the length of the scientific careers.

Categories and Subject Descriptors

H.3.5 [**Information Systems Applications**]: On-line Information Services; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Experimentation

Keywords

Scientometrics, DBLP, Collaboration Graph

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

KDD'12, August 12–16, 2012, Beijing, China.

1. 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, funding allocation, and so forth. With the growing of new on-line digital platforms, like DBLP, Google Scholar, Microsoft Academic Search, CiteSeer, etc., it is becoming easier than ever to explore and experiment ways of evaluating research products thanks to their aggregated information, like co-authorships, number of citations and related data. Despite this, even considering these web services, an author's publication and/or citation record gives only a partial account of the author's scientific profile. Indeed, in an evaluation process some co-authored works can unconditionally favor researchers who collaborated with those experts who were able to lead high-quality research projects. On the other hand, those who produced such relevant research products could not be distinguished (within their pure publication list) from their co-authors because of the typical assumption of a proportional collaboration among them.

Given these considerations, while many works evaluate the quantity and/or the quality of the research, (one among all, the h-index [2] and its different variations, a distinct part of the problem involves the study of co-authorship dependences, in terms of scientific influence among the authors. 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 DBLPVis, ArnetMiner or Microsoft Academic Search permit to visualize the academic career of a researcher through her co-authorship relations; however, they only show the relationships among authors based on the number of co-authored papers and are not suitable for such a challenging task.

In this paper we demonstrate a way to estimate this *dependence degree* among co-authors and present a set of graphical tools, integrated in the web environment *http://d-index.di.unito.it*, for permitting analyses, comparisons and evaluations of researchers' dependences.

2. BACKGROUND: D-INDEX

In this paper we make use of a novel measure of dependence between two co-authors, presented in [1], called d-index, which is calculated based on the *co-authorship network* that represents the common environment they have

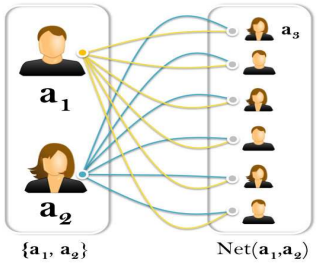

Figure 1: *Bipartite dependence Graph*: two authors a_1 and a_2 and the set of co-authors that they have in common (Net_{a_1,a_2}).

been working in. This is modeled through a *Bipartite Dependence Graph*, or *BDG*, which aim is to represent two authors in their scientific environment, defined as the set of co-authors they have in common. More in detail, given two authors a_i and a_j, we define their $BDG_{a_i,a_j} = \{V_{a_i,a_j}, E_{a_i,a_j}, w\}$, where

- $V_{a_i,a_j} = \{a_i, a_j\} \cup Net_{a_i,a_j}$, where Net_{a_i,a_j} is the set of researchers that co-authored at least one paper with both a_i and a_j;
- 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.

The weight w of each edge within the *BDG* represents the *relative dependence* of a collaboration between two authors with respect to a co-author. In other words, each weight represents how much the collaboration between a pair of authors is dependent on the collaboration with a common co-author. Considering the example shown in Figure 1, given the authors a_1 and a_2 and one of their co-authors, a_3, 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 the scientific assistance of a_2. This information quantifies how much the collaboration between a_1 and a_3 depends on the 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 $w^{a_j}_{a_i,a_k}$ of the collaboration between a_i and a_k on a_j as

$$w^{a_j}_{a_i,a_k} = 1 - \frac{p_{a_i,a_j,a_k} + p_{a_i,a_k,\neg a_j}}{p_{a_i,a_j,a_k} + p_{a_i,a_k,\neg a_j} + p_{a_j,a_k,\neg a_i}}, \quad (1)$$

where

- p_{a_i,a_j,a_k} is the productivity of the collaboration among a_i, a_j, and a_k calculated as the number of papers that a_i, a_j and a_k co-authored 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., calculated as the number of papers a_i and a_k co-authored without the assistance of a_j);
- $p_{a_j,a_k,\neg a_i}$ is the productivity of the collaboration between a_j and a_k without the contribution of a_i (i.e., calculated as the number of paper a_j and a_k co-authored without the assistance of a_i).

This relative dependence $w^{a_j}_{a_i,a_k}$ ranges from *0* to *1*; in particular, $w^{a_j}_{a_i,a_k} \approx 0$ indicates that the dependence of the scientific collaboration between a_i and a_k relatively to a_j is negligible, while a $w^{a_j}_{a_i,a_k} \approx 1$ highlights the contrary.

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 degree, *d*-index, of a_i on a_j by calculating the average dependence of all the scientific collaborations of a_i on the contribution of a_j. Thus, we calculate the dependence index, *d*-index, of a_i on a_j, as

$$d^{a_j}_{a_i} = \frac{p_{a_i,a_j}}{p_{a_i}} \times \frac{\sum_{a_k \in Net_{a_i,a_j}} w^{a_j}_{a_i,a_k}}{|Net_{a_i,a_j}|}, \quad (2)$$

where p_{a_i,a_j} is the productivity of a_i and a_j together (as before, calculated as the number of papers they co-authored together) and p_{a_i} is the total 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. The 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 on the co-authors can significantly differ, since they are based on their personal collaborations.

Now, given an author a_i and her list of *d*-index values related to her co-authors, we can also calculate her overall *dependence coefficient* as her dependence degree on the set of co-authors on whom she resulted more dependent. In a sense, we aim to quantify how much a researcher has been able to work without the scientists on whom she results more dependent. Thus, given an author a_i and the set of her co-authors on whom she resulted more dependent K_{a_i} (calculated applying an adaptive cut-off to the ordered list of d-index values), we calculate the dependence coefficient of a researcher a_i as her dependence degree on K_{a_i}. In particular, we first define the scientific network of a_i and K_{a_i} as $Net_{a_i,K_{a_i}} = Net_{a_i,a_1} \cup Net_{a_i,a_2} \cup \cdots \cup Net_{a_i,a_p}$, (where $K_{a_i} = \{a_1, a_2, \cdots a_p\}$). 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 authors in K_{a_i} as

$$w^{K_{a_i}}_{a_i,a_k} = 1 - \frac{p_{a_i,K_{a_i},a_k} + p_{a_i,a_k,\neg K_{a_i}}}{p_{a_i,K_{a_i},a_k} + p_{a_i,a_k,\neg K_{a_i}} + p_{K_{a_i},a_k,\neg a_i}}, \quad (3)$$

where

- p_{a_i,K_{a_i},a_k} is the number of paper co-authored by a_i, a_k, and at least one author in K_{a_i};
- $p_{a_i,a_k,\neg K_{a_i}}$ is the number of paper co-authored by a_i and a_k without the contribution of K_{a_i} (i.e., excluding the research outputs in which is also involved at least one author in K_{a_i});
- $p_{K_{a_i},a_k,\neg a_i}$ is the number of paper co-authored by a least one author in K_{a_i} and a_k without the contribution of a_i (i.e., excluding the papers 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

$$c_{a_i} = d^{K_{a_i}}_{a_i} = \frac{p_{a_i,K_{a_i}}}{p_{a_i}} \times \frac{\sum_{a_k \in Net_{a_i,K_{a_i}}} w^{K_{a_i}}_{a_i,a_k}}{|Net_{a_i,K_{a_i}}|}, \quad (4)$$

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}).

3. VISUALIZATION TOOLS

The web-based application, presented in Section 4, includes several tools for visualizing information about all the researchers indexed by the DBLP service. In this section we present them in detail.

3.1 Your and Others dependences

The first way to analyze the scientific dependences of a researcher is to plot the d-index values, in ascending order, related to her co-authors, as a curve within a Cartesian coordinate system (Figure 2(b)). This visualization scheme gives a first insight into an author's career by showing in a single chart the ordered dependences with respect to all her co-authors.

This curve can be also leveraged as an instrument for comparisons among different authors with respect to shared co-authors. However, considering that the co-authors of two researchers are most of the time different, in order to permit this comparison, we also present another 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 set of co-authors, $Net_{a_i} = \{a_{i,1}, a_{i,2} \cdots a_{i,n}\}$ and $Net_{a_j} = \{a_{j,1}, a_{j,2} \cdots a_{j,m}\}$ (where $a_{j,t}$ is the t-th co-author of a_j ordered by the d-index values, and n is not necessarily equals to m), we map onto the x-axis a set of meta-authors $Net_{meta} = \{a_{k,1}, a_{k,2} \cdots a_{k,r}\}$ where each author $a_{k,h} \in Net_{meta}$ represents one co-author of each 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 $r = n$.

This way, we permit to 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 on whom they are dependent. An example is shown in Figure 2(c).

3.2 You and your Collaborators

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, considering an author a_i, we introduce her *Mutual dependence Curve*, that plots all her co-authors onto the x-axis and, for each co-author a_k, it shows their *mutual dependence* calculated as $d^{a_k}_{a_i} - d^{a_i}_{a_k}$ onto the y-axis. The co-authors are ordered based on these values.

Using this simple tool, it is possible to distinguish the collaborators who depend on the considered researcher (those plotted under the x-axis) from those that more likely lead the scientific collaborations with the considered researcher (those plotted above the x-axis).

3.3 You and your Communities

Since researchers tend to work in communities, we provide a tool for analyzing scientific dependences among authors with respect to the communities they form. As the mutual dependence curve does, our tool called *Collaboration Map* shows the mutual dependences onto the y-axis. Regarding the x-axis, we aim to plot the co-authors based on the *local communities* they form with respect to the considered researcher. In particular, starting from an author a_i, we compute her square collaboration matrix C_i where both rows and columns represent her co-authors, and $C_i[j][k]$ is the number of papers of a_i co-authored by both a_j and a_k (the diagonal $C_i[j][j]$ simply reports the number of outcomes that a_i produced in collaboration with a_j). From this, we compute a dissimilarity matrix of co-authors D_i, where $D_i[j][k] = 1 - cos(C_i[j], C_i[k])$, 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 well-known X-Means clustering approach (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 (the communities are then visualized within the chart with different colors).

At this point, we need to map this information into a one-dimensional ordering (the x-axis). For this, without loss of generality, we use multi-dimensional scaling (MDS) proposed in [3] to embed the co-authors onto a 1-dimensional order. This chart allows us to graphically group the co-authors based on their scientific relations (i.e, the closer the authors onto the x-axis, the more productive is their collaboration with respect to the considered researcher), to identify the local communities in which the author is involved, and to estimate the roles the co-authors play within the local communities with respect to the considered researcher (for each co-author, the lower her y-value, the more scientifically dependent on the considered researcher). An example is shown in Figure 2(e).

3.4 You among All

Finally we provide a set of tools for comparing, from different points of view, the considered author with the all the scientists indexed in DBLP (1,097,418 researchers). In particular, we analyze, in three different charts, the relation between the dependence coefficient of an author (Section 2) and three parameters: the total number of her published papers, the total number of her co-authors and the length of her scientific career (expressed in years, from her first publication to her last one). Within these charts, all the researchers within DBLP are mapped with blue points, whereas the considered researcher is highlighted in red. This way, it is possible to compare each researcher with the entire community by focusing on different parameters, and according to the dependence coefficients. Moreover, within these charts, it is possible to map multiple authors and compare them with respect to the entire community.

4. DEMO SCENARIO

In this section we present a web environment, available at *http://d-index.di.unito.it*, for evaluation of scientific dependences among researchers. The considered data set contains information about the 1,097,418 authors and the 1,924,030 papers indexed by the DBLP bibliographic database[2]. Within this web platform it is possible to perform the following operations:

- visualization of the scientific profile of each researcher, including the information about her research papers and her collaborations over time;
- graphical analyses of scientific dependences with respect to the co-authors;

[2]Information updated on March 2012.

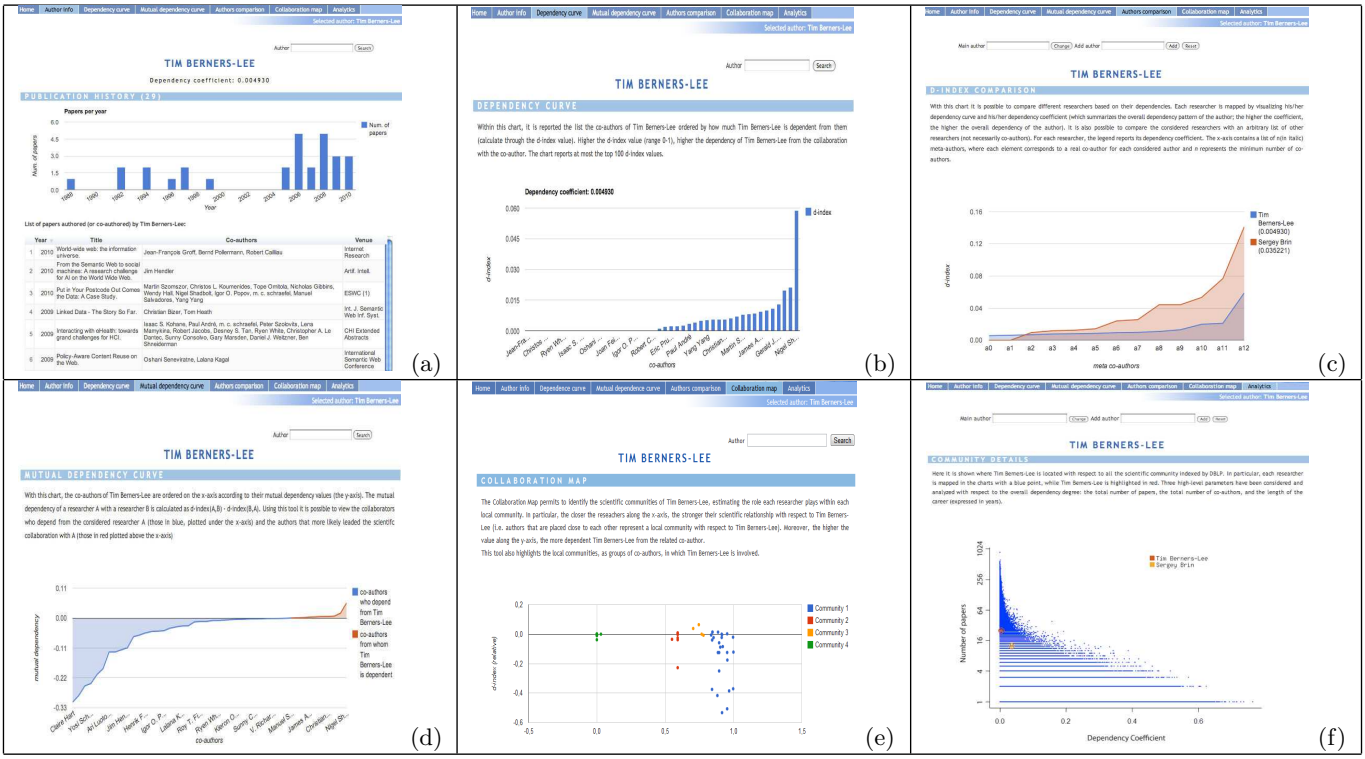

Figure 2: Different visualizations of the career of Dr. Tim Berners-Lee taken from *http://d-index.di.unito.it*; our web environment permits to visualize the information about the author (a), his dependence curve (here visualized as a histogram) (b), a comparison with other scientists (in this example only Dr. Sergey Brin)(c), his mutual dependences with respect to his co-authors (d), his local communities (e) and the relations between his dependence coefficient and his total number of papers compared with the whole scientific community (and, in this example, also focused on the comparison with Dr. Brin) (f).

- graphical comparisons among authors based on their dependence curves and coefficients;
- graphical analyses of the mutual dependences of the each researcher relatively to her co-authors;
- graphical analyses of the local communities in which the author is involved, estimating the dependence of each member on the considered researcher;
- 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 2 shows a set of screenshots taken from the presented web application. Within this system a user can search for authors and analyze their scientific profiles, according to the proposed measures. In particular, the system permits to query the database by typing an author's name and disambiguate, if it is necessary, the authors' list that match the query. The user can therefore analyze the scientific profile of the selected author (Figure 2(a)) by studying 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.), the entire list of co-authors over the entire career, and much more.

Then, as shown in Figure 2(b), it is possible to analyze her dependence curve that visualizes her scientific dependences (plotted as explained in Section 3.1). The curve permits to graphically order her co-authors with respect to their *d*-index value. The user can visualize this in two different ways (curve or histogram).

Using the dependence curve, it is also possible to compare the considered researcher with different authors; an example is shown in Figure 2(c). Then, the user can analyze her scientific relationships with her co-authors by visualizing her Mutual Dependences Curve (Figure 2(d)), graphically separating the authors who depend on the considered researcher (in blue) from those he is dependent on (in red). Moreover, by visualizing the Collaboration Map (Figure 2(e)), it is possible to get an alternative view of her dependences by also taking into account her communities (intended as groups of co-authors). In this way it is possible to analyze the research groups in which the author has been involved 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 (number of papers, number of co-authors and length of the career). Within these charts, it is possible to plot multiple authors (Figure 2(f)).

5. REFERENCES

[1] L. Di Caro, M. Cataldi, and C. Schifanella. The d-index: Discovering dependences among scientific collaborators from their bibliographic data records. *Scientometrics*, pages 1–25. 10.1007/s11192-012-0762-1.

[2] J. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 2005.

[3] W. S. Torgerson. *Theory and methods of scaling*. R.E. Krieger Pub. Co, 1958.