Tell Me Who Your Friends Are and I'll Tell You Who You Are: Studying the Evolution of Collaborations in Research Environments

Mario Cataldi, Myriam Lamolle
Université Paris 8
Paris, France
Email: m.cataldi,m.lamolle@iut.univ-paris8.fr

Luigi Di Caro, Claudio Schifanella
Università di Torino
Torino, Italy
Email: dicaro,schi@di.unito.it

***Abstract*—Nowadays, many tools and systems are available to allow the analysis and the comparison of researchers' scientific production. The reason underlying such interest is evident: promotions, funding allocations, and employments are currently based on the evaluation (and direct comparisons) of publication lists. Existing measures, like H-index, aim at supporting this process by automatic calculations of quality and/or quantity indices. In this work, we propose a demonstration of a web environment, available at *http://d-index.di.unito.it*, that faces the problem of studying the impact of collaborations in research communities. The presented system allows the estimation of the impact of each scientific collaboration on the production of each researchers indexed by the DBLP bibliographic database by means of a novel time-based modeling of collaborative environments. The proposed application provides several interactions and visualization schemes to deeply discover clear and latent insights, over time, around the work of each researcher. Furthermore, it allows cross-community rankings of the authors depending on similar collaboration patterns and dependences.**

I. Introduction

With the growth of new on-line digital platforms, like DBLP[1], Google Scholar[2], Microsoft Academic Search[3], CiteSeer[4], etc., it is becoming easier than ever to explore and experiment ways of evaluating research activities. These tools permit to aggregate information, like co-authorships and number of citations, comparing authors (also called scientists, or researchers, from now on) and/or research products based on different existing metrics [1], [2], [3].

However, even considering these environments, the author's publication and/or citation record gives only a partial account of the author's scientific profile. In fact, the outcome of an author is usually the result of multiple scientific collaborations, which analysis is usually not taken into account by the existing systems. Most of them commonly assume the co-authorship to be a proportional collaboration among the involved authors, missing out their relationships and their relative scientific impact on the resulting work [4]. Based on the same assumptions, [5] stated that an author's scientific relevance should not be based on a pure quantification of his/her publication history, but it should be evaluated based on how much co-workers he/she has been able to connect to in order to produce (joint) scientific publications.

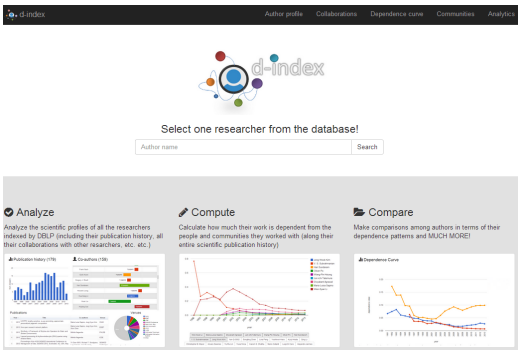

Fig. 1. The homepage of *http://d-index.di.unito.it*

Indeed, a research collaboration can be defined as a two-way process where individuals and/or organizations share learning, ideas and experiences to produce scientific outcomes. Collaborations are obviously necessary, because of the evident difficulty for individual scientists to conduct several groundbreaking research on their own. For this reason, one of the key (and more demanded) aspect of a successful researcher is the development of a large network of collaborators that can help the researcher bring new solutions and propose, continuously, novel ideas and approaches to the research community. On the other hand, the evaluation of individual researchers needs a sort of inverse process with the primary goal of understanding the role of each researcher and his/her specific impact, on the research community, in this collaborative environment.

According to these ideas, as also stated by [6], [7], it is crucial to study the scientific relationships 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. In fact, in an evaluation process, some co-authored works can unconditionally favor researchers who worked in a bigger research environment that involved experts able to lead high-quality research projects, while, on the other hand, those who produced such relevant research products could not be distinguished (from the pure publication history) from their co-authors because of the typical assumption of a proportional collaboration among them.

With these goals in mind, based on the entire DBLP bibliographic database, in this paper we demonstrate a web platform

[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*

(available at *http://d-index.di.unito.it*), which implements a novel temporal model that aims at studying the collaborations of an author and estimating their impacts, over time, on his/her scientific production. This system first allows the user to study the scientific profile of each researcher and analyze, through several dynamic visualization tools, the evolution of the impact of each collaboration on his/her scientific output. The user can also study and compare the overall collaboration patterns of the researchers (expressing how much each scientist results dependent, over time, on the collaboration with the surrounding community) and rank, by using different parameters, the performance of each author by focusing on his/her collaboration history. By using these tools, it is therefore possible to detect anomalies with respect to average patterns of similar authors, and compare each career, over time, with respect to all the researchers indexed by DBLP.

II. Analyzing Scientific Collaborations

In this work we present a web platform that permits to study and visualize the impact, over time, of each scientific collaboration on the production of a researcher. This measure will permit to estimate the independence of an author from the collaboration with other scientists and his/her potential capacity to maintain the same production in a different collaborative environment.

With this goal, given an author a_i, we first formalize his/her set of research outputs (also called papers, works or outcomes) $O^t_{a_i} = \{o^t_{a_i,1}, o_{a^t_i,2}, \cdots o^t_{a_i,n}\}$, as the set of scientific outcomes authored, or co-authored by a_i at the time t[5]. Considering this information, it is possible to quantify the "productivity" of a_i, $p^t_{a_i} = |O^t_{a_i}|$, at time t, as the cardinality of $O^t_{a_i}$. This approach is extendable to any set of authors with any cardinality.

Moreover, given an author a_i, we formalize the scientific network $Net^t_{a_i} = \{a^t_1, a^t_2, \cdots, a^t_h\}$ in which he/she produced research outcomes as the set of his co-authors, at time t, of a_i. In the same way, given two authors a_i and a_j, we formalize their common scientific collaboration network $Net^t_{a_i,a_j}$, as the set of authors who co-authored, at time t, at least one paper with both a_i and a_j.

Following these basic formalizations, we first aim at studying each scientific collaboration of an author and estimate its autonomy from the surrounding scientific community. Then, we will quantify the overall dependence, called d-index, of a considered researcher on the scientific collaboration with a specific co-author by analyzing how much each collaboration of the first is independent of the contribution of the second. Based on these high level assumptions, we first introduce a method to measure the "autonomy" of a collaboration, by taking into account the common scientific production, at time t, of the involved authors. In this respect, given two authors a_i, a_j and their common scientific network $Net^t_{a_i,a_j}$, the autonomy of their collaboration $w^t_{a_i,a_j}$ is calculated as

$$w^t_{a_i,a_j} = \begin{cases} 0 & \text{if } Net^t_{a_i,a_j} = \{\} \\ \frac{1}{\sum_{a_k \in Net^t_{a_i,a_j}} \left(\sum_{x=1}^{collab(a_k, O^t_{a_i,a_j})} \frac{1}{x} \right)} & \text{if } Net^t_{a_i,a_j} \neq \{\} \end{cases}$$

where the function $collab(a_k, O^t_{a_i,a_j})$ returns the number of times the author a_k co-authored a paper with both a_i and a_j at time t. This formula permits to measure the independence, at time t, of the collaboration between a_i and a_j from the collaboration with any other author of the common scientific environment, expressed by $Net^t_{a_i,a_j}$. In this way, we take into account number and frequency of each collaboration; in fact, from one side, we count how many external co-authors, along their collaboration history (until t) have been involved in the collaboration between a_i and a_j. From the other side, we also aim at evaluating the frequency of each contribution on their collaborations. In a sense, the autonomy of the collaboration will be lower when a high number of external co-authors are repetitively involved in the scientific outputs of the collaboration. The higher the autonomy, the more independent the work of a_i and a_j of the collaboration with any other co-author (and the other way around). From this, given an author a_i, we aim at estimating his/her overall dependence, called d-index, on the collaboration with a co-author a_j by taking into account the capacity of a_i of working in his/her scientific environment without the scientific support of a_j. For this, given an author a_i and his/her scientific environment $Net^t_{a_i}$, at time t, we define the dependence value, d-index, of the co-author a_i on the collaboration with a_j as $d^t_{a_i \to a_j}$

$$d^t_{a_i \to a_j} = \frac{p^t_{a_i,a_j}}{p^t_{a_i}} \times \frac{w^t_{a_i,a_j,Net^t_{a_i}} + w^t_{a_j,\neg a_i,Net^t_{a_i}}}{w^t_{a_i,a_j,Net^t_{a_i}} + w^t_{a_j,\neg a_i,Net^t_{a_i}} + w^t_{a_i,\neg a_j,Net^t_{a_i}}},$$

where

- $p^t_{a_i}$ returns the productivity of a_i at time t;
- $p^t_{a_i,a_j}$ is the productivity of the collaboration between a_i, a_j at time t;
- $w^t_{a_i,a_j,Net^t_{a_i}}$ is the autonomy score of the collaboration, at time t, among a_i, a_j and $Net^t_{a_i}$ (i.e. the autonomy score of the collaboration between a_i and a_j, and at least one author a_k in $Net^t_{a_i}$);
- $w^t_{a_i,\neg a_j,Net^t_{a_i}}$ is the autonomy score of the collaboration between at least one author in $Net^t_{a_i}$ and a_i without the contribution of a_j (i.e., excluding the research outputs in which a_j is also involved);
- $w^t_{a_j,\neg a_i,Net^t_{a_i}}$ is the autonomy score of the collaboration between a least one author in $Net^t_{a_i}$ and a_j without the contribution of a_i (i.e., excluding the research outputs in which a_i is also involved);

It is important to notice that the d-index value $d^t_{a_i \to a_j}$ ranges from 0 to 1; in particular, $d^t_{a_i \to a_j} \approx 0$ indicates that the dependence of a_i on a_j, at time t, is negligible, while a $d^t_{a_i \to a_j} \approx 1$ highlights the contrary. In fact, the second term of the formula increases when the autonomy score of a_i and $Net^t_{a_i}$, without the contribution of a_j, is negligible ($w^t_{a_i,\neg a_j,Net^t_{a_i}} \approx 0$) and the other collaborations are significantly autonomous ($w^t_{a_i,a_j,Net^t_{a_i}} > 0$ and $w^t_{a_i,\neg a_j,Net^t_{a_i}} > 0$). On the other hand, the higher the $w^t_{a_i,\neg a_j,Net^t_{a_i}}$, the lower the relative dependence.

Please also notice that $d^t_{a_i \to a_j} \neq d^t_{a_j \to a_i}$; in fact, their mutual dependences can significantly differ, since they are also based on their personal collaborations (which are obviously not the same, even if they share the same co-authors).

[5] In this paper, the considered time intervals represent years.

A. The Dependence Trajectory: Visualizing the Evolution of the Dependence on the Surrounding Community over Time

In the previous Section, we introduced a novel way to estimate the dependences, at a specific time, of a given author on the scientific collaborations with all his/her co-authors. These values can be now leveraged to graphically map the scientific dependences of an author, along his/her career, on the collaboration with each co-author, as a set of curves that plots the relative d-index values. For this, we define the dependence curve of an author a_i with respect to a co-author a_j as

$$\overrightarrow{d_{a_i \to a_j}} = \{d^t_{a_i \to a_j}, d^{t+1}_{a_i \to a_j}, \cdots, d^{t+n}_{a_i \to a_j}\},$$

where t is the year of the first publication of a_i, and n is the career of a_i expressed as the arithmetical difference between the last and the first year of publication of a_i. Thus, given an author a_i, and the complete set of his/her co-authors expressed by Net_{a_i}, it is now possible to graphically represent, in the same chart, his/her dependences on each co-author $a_k \in Net_{a_i}$, along the career of a_i, to obtain a first sight on this mined knowledge. Each of these curves can graphically highlight the evolution of the collaboration with the related co-author over time and permits to understand how much the considered author became independent (or dependent) of him/her with the years. An example is shown in Figure 2: the evolution of a subset of the collaborations of Dr. Jure Leskovec[6] is visualized. In this example it is clearly visible how, for example, the considered scientist was increasingly dependent on the collaboration with Dr. Christos Faloutsos until 2008 (Ph.D. supervisor until 2008). Then, after that moment, the work of Dr. Leskovec started becoming independent of this scientific collaboration due to his change of position and work conditions (i.e., the dependence curve related to Dr. Faloutsos starts lowering after 2008). Thus, the curve permits to graphically summarize the evolution of the scientific relationship with a co-author that could be hardly identifiable from the pure publication history (please notice that, even if both scientists continued collaborating after 2008, the presented measure is able to identify a drastic change in their collaborations pattern).

Despite that, we also aim at obtaining a one-curve evaluation system to summarize, at best, the overall independence of the author. The high-level goal is to graphically highlight the evolution of the average dependence of the author on the collaboration with others. In a sense, the curve permits to study how the career of an author evolved with respect to the dependence on the collaboration with his/her surrounding scientific community.

Thus, given the complete set of dependence curves of a researcher, we calculate his/her *dependence trajectory*, by calculating the average standard deviation, along the time, of each d-index value, for each co-author, from the optimal attended value of 0 (which would mean a dependence score of 0; i.e., the production of the considered author is independent of the collaboration with the considered co-author). More

[6]Dr. Jure Leskovec is a scientist in the social network analysis community. He is currently Assistant Professor of Computer Science at Stanford University. He completed his Ph.D. program under the supervision of Dr. Christos Faloutsos in September 2008.

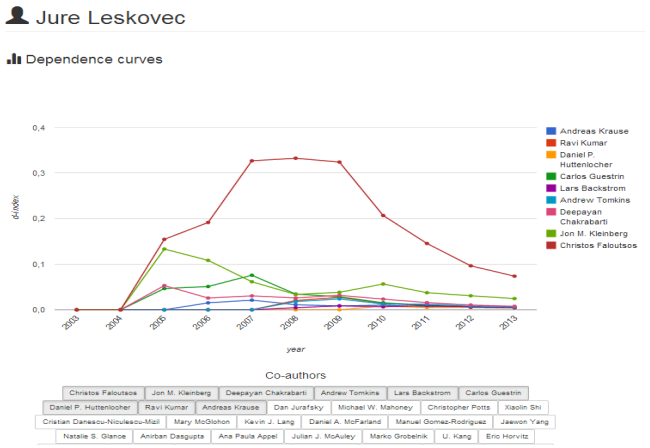

Fig. 2. Evolution of dependences of Dr. Jure Leskovec with respect to a selected subset of his co-authors.

formally, given an author a_i, we define his/her dependence trajectory $\overrightarrow{d_{a_i}}$ as

$$\overrightarrow{d_{a_i}} = \{sd^t_{a_i}, sd^{t+1}_{a_i}, \cdots, sd^{t+n}_{a_i}\},$$

where $sd^t_{a_i}$ is calculated as

$$sd^t_{a_i} = \sqrt{\frac{\sum_{a_k \in Net_{a_i}} (d^t_{a_i \to a_k})^2}{|Net_{a_i}|}}.$$

The meaning of this formula is simple: it calculates the average standard deviation of the previously calculated d-index values from the optimal value of 0. The higher the $sd^t_{a_i}$, the more dependent the work of a_i on the collaboration with the surrounding community (expressed by the complete set of his/her co-authors) at time t.

Notice that, using the dependence trajectory, we can now compare researchers focusing on their average collaboration behavior, and estimate how much the productivity of each individual researcher depends, on average, on the interactions with the scientific environment in which he/she has been involved until that time. In a sense, through the dependence trajectory, we analyze the impact of the surrounding environment, along the time, on the publication record of the considered researcher. Moreover, it is now possible to rank the indexed authors based on their independence of their environment and compare their collaboration patterns with respect to the whole community and/or sub-communities of similar authors.

III. Application and Demo Scenario

In this section we present our application, available at *http://d-index.di.unito.it*, for the analysis and comparison of scientific collaboration patterns of researchers. For this, we considered the DBLP data set[7], containing information about 1,342,723 authors and 2,446,236 scientific papers [8]. Within this web platform it is possible to perform the following operations:

- visualize the scientific profile of each researcher, including the information about the published research

[7]http://dblp.uni-trier.de/db

[8]Information updated on March 2014.

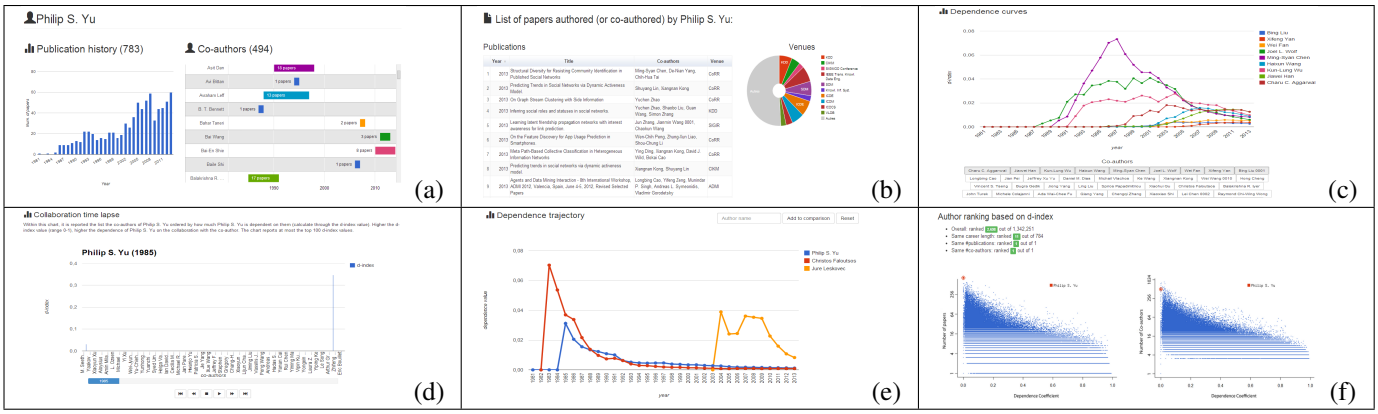

Fig. 3. Different visualizations, taken from *http://d-index.di.unito.it*, of the profile and collaborations of Philip S. Yu (one of the most prolific author on DBLP). The web environment permits to analyze the author profile (a) and (b), the evolution of the collaboration patterns with his co-authors (c), the "time-lapse" of his scientific dependences (d), his dependence trajectory (here also compared with the ones of Dr. Faloutsos and Dr. Leskovec) (e) and rank and compare his behavior with respect to the rest of the scientific community (f).

outcomes, venues and the quantification of each scientific collaborations over time;

- visualize the "time-lapse" of the evolution of the scientific dependence of an author with respect to all his/her co-authors;
- compare authors based on their dependence trajectories;
- compare the dependence scores of an author with the ones of similar scientists (i.e., same career length, same number of published papers, same number of co-authors, etc.);
- rank, based on different parameters, any author with respect to the entire DBLP community and/or sub-communities formed by similar authors.

Figure 3 shows a set of screen-shots taken from the presented web application. Within this system the user can search for any author indexed by DBLP and analyze his/her scientific profile, focusing on his/her scientific collaborations over time. The system allows the user to query the database and disambiguate, when necessary, the authors list that, eventually, partially matches the query. The user can therefore analyze the profile of the selected author by studying his/her scientific career (visualized as a histogram of papers per year), his/her list of scientific outcomes (through the standard information as the title, the name of the conference/journal, the year of the publication, venue, etc.), the collaboration history with each co-author over time (though a chart visualizing the distribution of co-authored outcomes along the career), the diagram representing the set of publication venues, and other insights (Figures 3(a) and (b)) .

Then, as shown in Figure 3(c), it is possible to analyze the complete set of scientific collaborations of an author and their evolution and impact, over time, on the work of the considered author. The dependence on the collaboration with each co-author is visualized as a curve, mapping the evolution of the dependence (d-index) of the author on the support of each co-author along the career of the former. Within this chart it is possible to select/deselect the desired co-authors to better analyze and compare their collaborations over time. An interactive visualization, called "time-lapse" allows the user better focusing on a specific time interval and/or a subset of co-authors (Figure 3(d)).

Moreover, it is also possible to analyze the overall independence of an author from the surrounding community through the visualization of his/her dependence trajectory (Figure 3(e)). This curve permits to graphically visualize the evolution of the career of an author by focusing on how much his/her overall work can be considered dependent on the interactions with the surrounding community. In this chart it is also possible to compare the author against others researchers (even if they do not share the same time career) and/or analyze his/her behavior with the respect to the average dependence pattern of the authors with the same career length (i.e., it is possible to graphically analyze how the author performed with respect to those who worked during the same amount of time).

Finally, the application permits to compare and rank each researcher with the entire DBLP community with respect to different parameters (number of papers, number of co-authors and length of the career). Within these charts, it is possible to rank the selected author with respect to the whole community and/or the subset of authors with similar characteristics (same career length, same number of papers/co-authors). An example is shown in Figure 3(f).

References

[1] A. Pepe and M. Rodriguez, "An in-depth longitudinal analysis of mixing patterns in a small scientific collaboration network," *Scientometrics*, vol. 85, no. 3, 2010.

[2] M. Rodriguez and A. Pepe, "On the relationship between the structural and socioacademic communities of a coauthorship network," *Journal of Informetrics*, vol. 2, no. 3, pp. 195–201, 2008.

[3] C. Taramasco, J. Cointet, and C. Roth, "Academic team formation as evolving hypergraphs," *Scientometrics*, vol. 85, no. 3, pp. 721–740, 2010.

[4] R. M. Slone, "Coauthors' contributions to major papers published in the ajr: frequency of undeserved coauthorship." *AJR. American journal of roentgenology*, vol. 167, no. 3, pp. 571–579, 1996.

[5] M. Ausloos, "A scientometrics law about co-authors and their ranking: the co-author core," *Scientometrics*, vol. 95, no. 3, pp. 895–909, 2013.

[6] P. Van den Besselaar, U. Sandström, and I. Van der Weijden, "The independence indicator: Towards bibliometric quality indicators at the individual level," 2012.

[7] L. D. Caro, M. Cataldi, and C. Schifanella, "The d-index: Discovering dependences among scientific collaborators from their bibliographic data records," *Scientometrics*, vol. 93, no. 3, pp. 583–607, 2012.