Co-Citations and Relevance of Authors and Author Groups

Maria Bras-Amorós, Josep Domingo-Ferrer and Albert Vico-Oton

Universitat Rovira i Virgili
Dept. of Computer Engineering and Mathematics
UNESCO Chair in Data Privacy
Av. Països Catalans 26
E-43007 Tarragona, Catalonia
E-mail {maria.bras,josep.domingo,albert.vico}@urv.cat

Abstract

The way an author or a group of authors are cited tells more about the real impact of their work than authorship and collaborations. Indeed, the connections within the scientific community can be more accurately elicited from the co-citation graph than from the collaboration graph. We suggest some indices that can be drawn from the co-citation graph in order to capture the relevance of individual authors and the relevance of groups of authors.
Keywords: Co-citation graph, collaboration graph, bibliometry, relevance indices.

1 Introduction

Usually the proximity or mutual influence between authors is investigated in terms of the collaboration graph, in which each author is represented by a node and two authors are connected if they have co-authored at least one paper. The collaboration graph has attracted some attention from the mathematical community. For instance we can find collaboration graphs in the web pages of several mathematics departments, like the ones at the University of Georgia (Figure 1), Oakland University (Figure 2), or the Naval Postgraduate School in Monterey, California (Figure 3).

The Erdös number, reflecting the collaboration distance between a certain author and the mathematician Paul Erdös (who directly collaborated with 511 authors in his lifetime), is a popular index computed on the collaboration graph. Professor Jerrold W. Grossman leads a project devoted to the Erdös number [4]. More generally, the collaboration graph is used to find the collaboration distance between two mathematical authors. The Erdös number of mathematicians and collaboration distances between any two mathematical authors can be

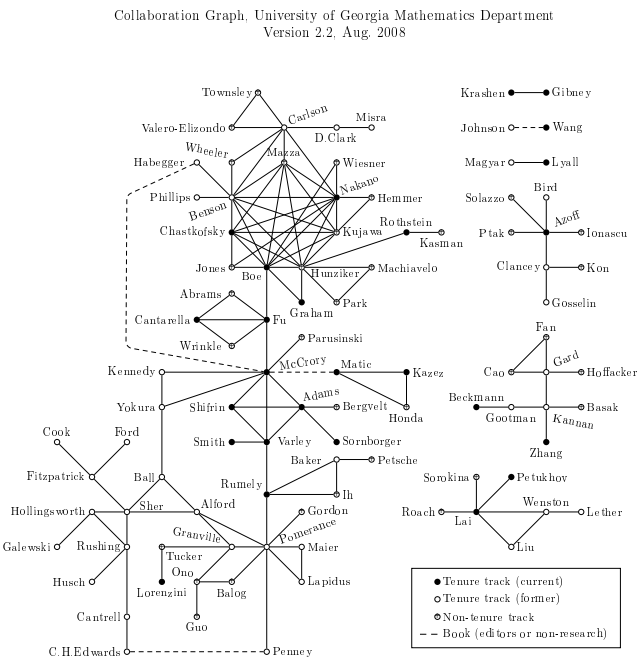


Figure 1: Collaboration graph of the Mathematics Department at the University of Georgia

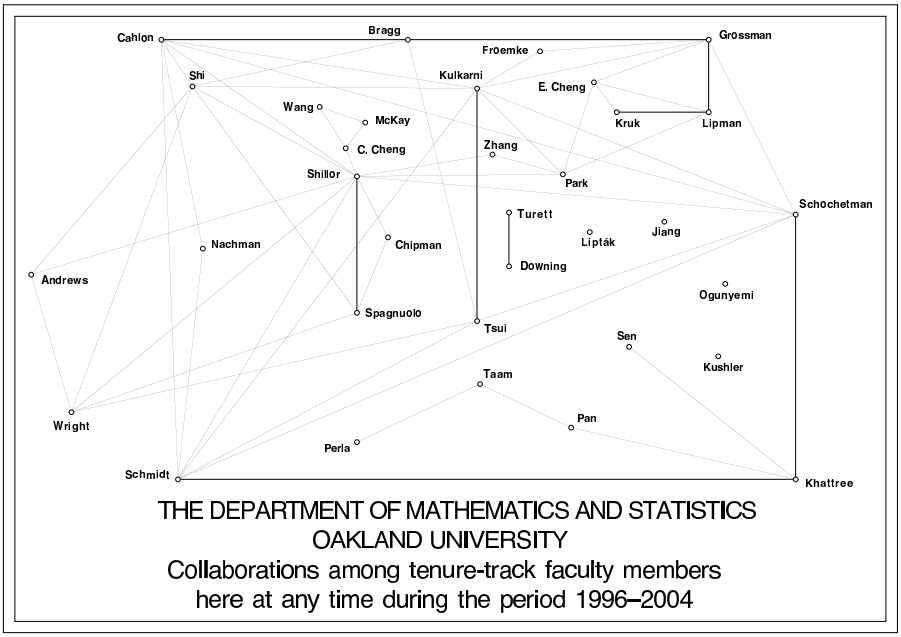


Figure 2: Collaboration graph of the Mathematics Department at Oakland University

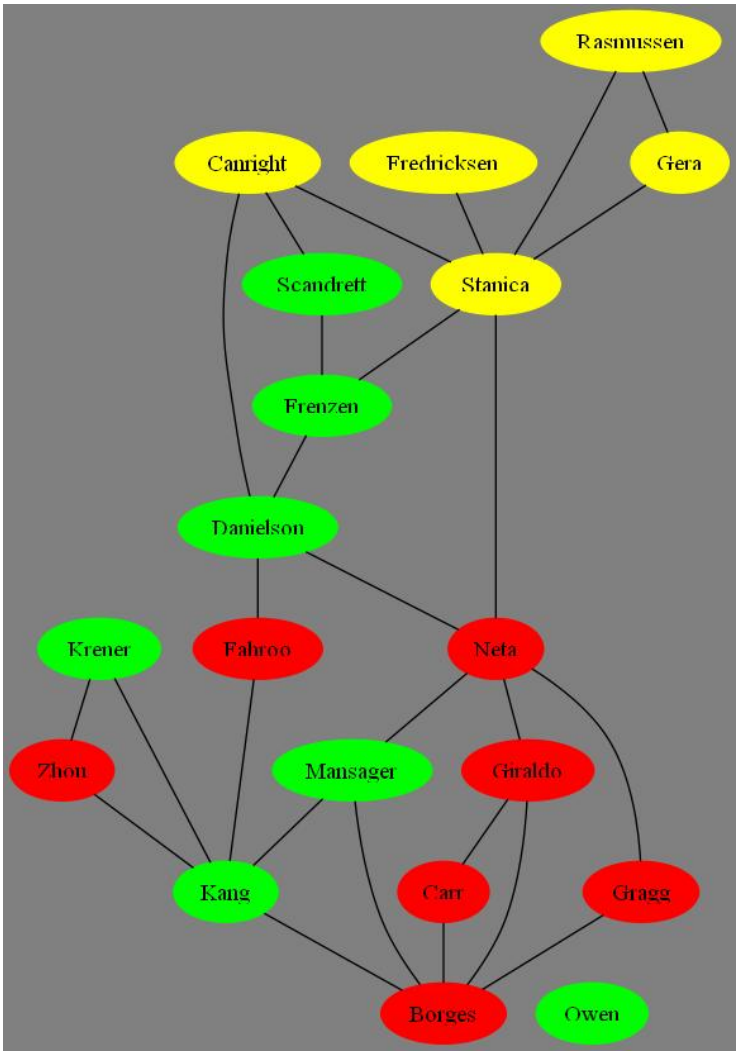


Figure 3: Collaboration graphs of the Mathematics Department at the Naval Postgraduate School in Monterey

automatically computed using the MathSciNet database run by the American Mathematical Society [14].

Scientific studies about the collaboration network can be found, for instance in Grossman's contributions [5, 6], in Mark E. J. Newman's contributions [9, 10, 11, 12], or in [2] by Batagelj and Mrvar. It is quite standard to use a more accurate version of the collaboration graph, the weighted collaboration graph, where the edges are weighted according to the number of collaborations.

Rather than the collaboration graph, we will use in this paper the co-citation graph, in which the nodes corresponding to two authors are connected by an edge if there exists a paper simultaneously citing both authors. The co-citation concept was first defined by Henry Small in [13]. This author defined the co-citation between two papers as the frequency with which two items of earlier literature are cited together by later literature. We resume that co-citation idea but we apply it to the authorship of papers, that is, we focus on the relationships between authors rather than papers.

There are some contributions in the literature using co-citations to measure author features. In [1], Ahlgren, Jarneving and Rousseau classify authors by indices that try to establish their similarity in view of clustering them. In [8], concepts of information theory (like mutual information) are applied with a similar aim of creating clusters of authors.

While Small focuses on the concept of a "core" paper, the one with most co-citations among the papers in a certain subject, we use co-citations to derive indices to measure the relevance of authors and groups of authors.

Contribution and plan of this paper

It is widely accepted that citations tell at least as much about an author as authorship itself does. This is the origin, for instance, of the h-index [7]. We propose to use co-citations to derive indices measuring the relevance of authors and groups of authors.

In Section 2 we formally define co-citations and the co-citation graph. In Section 3 we use the co-citation graph to define three indices of relevance for individual authors. In Section 4 we distinguish between the relevance of an author and the relevance of a group of authors, and we give several indices based on co-citations that measure the relevance of a group of authors, in the sense of evaluating how present the group is in the citations by papers in a certain subject. Section 5 contains some conclusions.

For the sake of realism, in the examples throughout this paper we mention real names of authors and departments whose bibliometric data are publicly available. We wish to emphasize that it is not our purpose to judge such authors or departments in any way. In fact, our examples are unlikely to give an accurate portrait of the mentioned authors or departments, for at least two reasons: i) no single bibliometric database is guaranteed to index 100% of the scientific production or the citations of anybody; ii) our examples may not be current because they are based on the database contents at the time of writing the first version of this paper.

2 The co-citation graph

Co-citations and the co-citation graph can be defined as follows.

Definition 1 (Co-citation)**.** *Let $G(p)$ be the set of authors cited in the list of references of a paper p. Paper p is a co-citation to an unordered pair of authors (i, j), with $i \neq j$ if $i, j \in G(p)$. In plain words, a paper simultaneously citing an unordered pair of distinct authors is a co-citation to that pair of authors.*

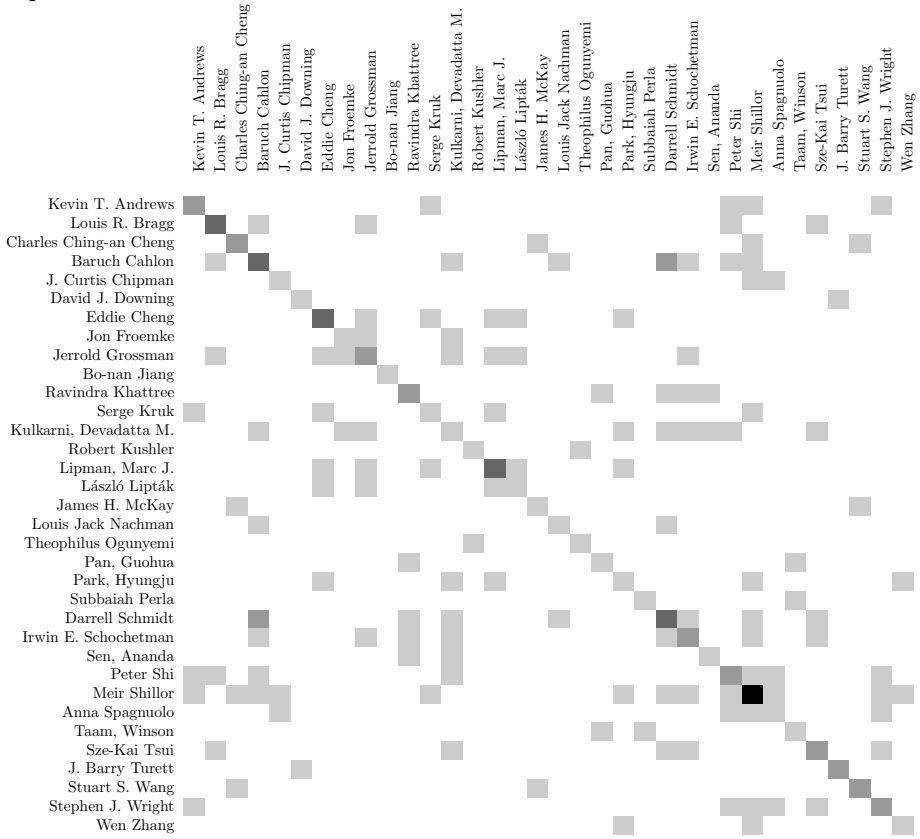
Definition 2 (Co-citation graph)**.** *The* co-citation graph *is a graph having authors as nodes such that an undirected edge between two authors i and j exists if at least one co-citation to the pair (i, j) exists. The* weighted co-citation graph *is a co-citation graph where each edge is assigned a weight equal to the number of co-citations to the pair of authors (i, j) connected by the edge.*

Our hypothesis is that, while the connections in the collaboration graph show the mathematical social network, the connections in the co-citation graph show better the connections in terms of visibility or impact of the published work. For instance, two names mentioned in Section 1, Jerrold W. Grossman and Mark E. J. Newman, corresponded to non-adjacent nodes in the collaboration graph at the time of collecting data for this paper (according to the MathSciNet database), but they are connected in the co-citation graph, because they are often cited together.

In Table 1 we show the collaborations between authors in the Mathematics Department of Oakland University. A white cell means no collaboration (no edge in the collaboration graph) and a black cell means maximum number of collaborations, which in this case is 132 and is the number of publications of Professor Meir Shillor. It is important to notice that the highest number of collaborations of any author is with herself/himself, and equals the number of her/his publications. In the off-diagonal cells of Table 2 we show the co-citations between authors in the same department. Cells along the diagonal represent the number of papers which contain citations to each author (citing papers); note that the number of citing papers to an author X may be less than the number of citations to X, because a citing paper can cite several papers by X. A white cell means no co-citation/citing paper and a black cell means maximum number of co-citations/citing papers. In this case, the black cell is 233 and corresponds to the number of papers citing Professor Meir Shillor. Again, the highest values for each author are in the diagonal: by definition, the number of co-citations of an author X with any other author can be no more than the number of papers citing X. Note that:

- Two authors i and j can have collaborations and no co-citations: this happens when no single paper cites their collaborations and, in addition, no single paper cites independent publications by i and by j. For example, Table 1 shows that, when the data were collected, Eddie Cheng and Serge Kruk had collaborated (they actually had 5 publications in common) but Table 2 shows that they had no co-citations.

Table 1: Number of collaborations between authors of the Oakland Mathematics Department



- Two authors i and j can have co-citations and no collaborations: this happens when at least one paper cites at least one paper by i and at least one paper by j. The above mentioned example of Jerrold W. Grossman and Mark E. J. Newman illustrates a case of co-citations without collaborations.

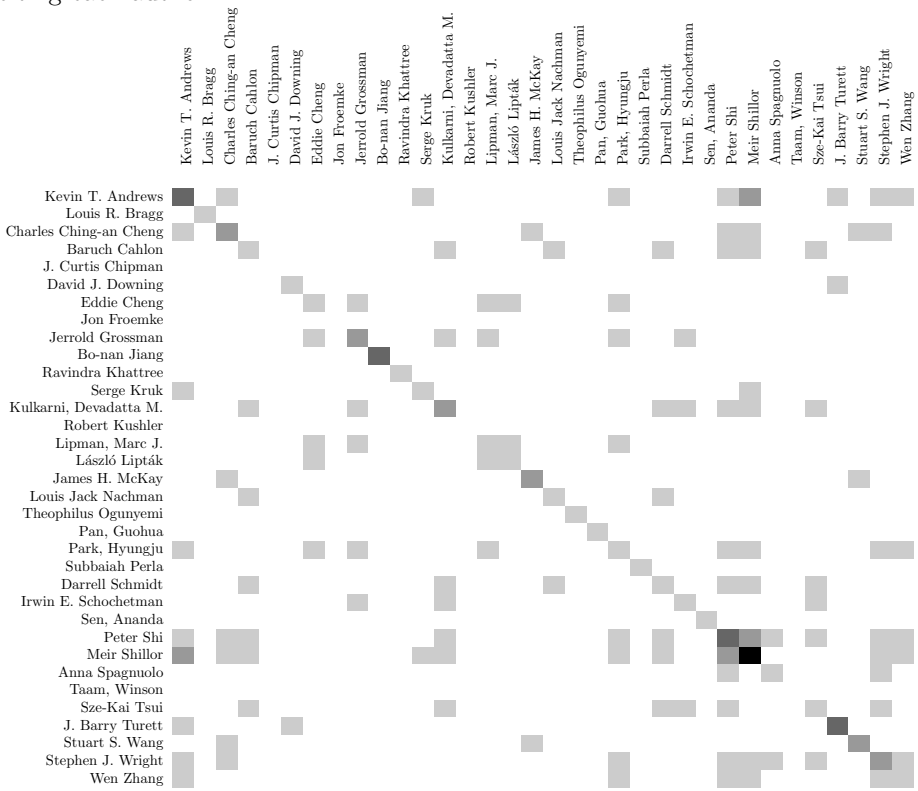
3 Relevance of individual authors

We suggest three indices for the relevance of an individual author.

Maximum co-cited count

This index counts the number of authors for which a given author is maximum co-cited. A formal definition of the indicator follows.

Table 2: Number of co-citations between authors of the Oakland Mathematics Department (off-diagonal cells). Diagonal cells represent the number of papers citing each author.



Definition 3. *The* maximum co-cited count *of author i is m if and only if there exists a set of authors $\{i_1, \cdots, i_m\}$, where $i_k \neq i$ for all $1 \leq k \leq m$, such that the following conditions hold:*

- *Author i is co-cited with each author in the set $\{i_1, \cdots, i_m\}$;*
- *For all $1 \leq k \leq m$, in the weighted co-citation graph the edge (i, i_k) has maximum weight among those edges incident to i_k;*
- *The previous condition does not hold for any other authors not belonging to $\{i_1, \cdots, i_m\}$*

Example 1. Assume that author i is co-cited 4 times with author i_1, 5 times with author i_2 and 6 times with author i_3. At the same time, i_1 is not co-cited more than 4 times with any other author different from i, i_2 is not co-cited more than 5 times with any other author different from i and i_3 is not co-cited more than 6 times with any other author different from i. That is, i is the maximum co-cited author for authors in the set $\{i_1, i_2, i_3\}$. If there is no other author for whom i is the maximum co-cited author (last condition of Definition 3), then the maximum co-cited count of author i is 3.

Weighted maximum co-cited count

This index counts the number of authors to which a given author is maximum co-cited weighted by the number of co-citations. A formal definition follows.

Definition 4. *The* weighted maximum co-cited count *of author i is $\sum_{k=1}^{m} c(i, i_k)$ if and only if there exists a set of authors $\{i_1, \cdots, i_m\}$, where $i_k \neq i$ for all $1 \leq k \leq m$, such that the following conditions hold:*

- *Author i is co-cited with each author in the set $\{i_1, \cdots, i_m\}$;*
- *For all $1 \leq k \leq m$, in the weighted co-citation graph the edge (i, i_k) has maximum weight among those edges incident to i_k;*
- *The previous condition does not hold for any other authors not belonging to $\{i_1, \cdots, i_m\}$;*
- *The weight of (i, i_k) is $c(i, i_k)$.*

Co-citation entropy

The co-citation entropy measures how transversal an author is perceived by the community. For example, let there be two authors i_1 and i_2 who have the same number of co-citations (sum of weights of edges incident to each author in the weighted co-citation graph), but author i_1 has been co-cited with a very small set of other authors, whereas author i_2 has been co-cited with a larger set of authors. In this case, the work of i_1 is perceived as "clique work", whereas the work of i_2 is seen as more transversal, because i_2 is cited together with a higher diversity of other authors. A formal definition of the co-citation entropy follows.

Definition 5. *Given an author i, let $w_1(i), \cdots, w_{n_i}(i)$ be the weights of the edges incident to the node of that author in the weighted co-citation graph. Define $W(i)$ as the sum of the previous weights and the relative weights as $rw_k(i) = w_k(i)/W(i)$, for $k = 1$ to n_i. Consider the set of relative weights as a probability distribution over the authors $i_1, \cdots, i_{n_i}$ co-cited with i. The Shannon entropy $H(\{rw_1(i), \cdots, rw_{n_i}(i)\})$ of that distribution is the co-citation entropy of i.*

The greater the co-citation entropy of an author, the more scattered and evenly distributed are her/his co-citations, and the greater is her/his transversality.

In Table 3 we show the above relevance indices computed for all authors in the Oakland Mathematics Department, with the authors sorted by increasing order of the weighted maximum co-cited count. One can appreciate some correlation between the three indices.

4 Relevance of a group of authors

Just like the relevance of an author can be measured by how present is she/he in the citations of a certain subject matter, the relevance of a group of authors can be measured by how present is the group of authors in those citations. A group of authors is understood here in a broad sense, and includes research groups, research institutes, university departments, entire universities and even all scientists in a country.

There are several conceivable group relevance metrics. We will focus on metrics that are monotonic with set inclusion, that is, such that, if A, B are groups of authors with $A \subseteq B$, then the relevance of A is less than or equal to the relevance of B. In other words, we consider metrics such that adding new authors to a group does not decrease the relevance of the group. Subadditivity is another reasonable property of a group relevance metric: if A and B are two groups of authors, the relevance of $A \cup B$ is no more than the sum of the relevance of A plus the relevance of B. Even if $A \cap B = \emptyset$, the relevance of $A \cup B$ can be less than the sum of the relevances of A and B; in particular, the relevance of a group of authors can be less than the sum of the relevances of individual authors in the group. Consider the following two extreme examples.

Example 2. In a group in which all members of the group sign all papers written by anyone in the group (such groups exist in real life!), the citations received by the group are exactly the citations received by any individual member of the group. Hence, if we use a relevance metric proportional to received citations, the relevance of the group is the same as the relevance of any individual member.

Example 3. In a group where all members publish independently and no paper is ever coauthored between two group members, the citations received by the group are the result of adding the citations received by its members. Hence, if we use a relevance metric proportional to received citations, the relevance of the group is the sum of the relevances of its individual members.

Table 3: Relevance indices of authors of the Oakland Mathematics Department sorted by increasing order of the weighted maximum co-cited count

Author	Maximum co-cited count	Weighted maximum co-cited count	Co-citation entropy
Subbaiah Perla	0	0	0
Jon Froemke	0	0	0
Winson Taam	0	0	0
J. Curtis Chipman	0	0	0
Robert Kushler	0	0	0
Wen Zhang	6	7	5.8287
Theophilus Ogunyemi	14	14	4.3219
Ananda Sen	16	16	4.6438
Guohua Pan	42	47	5.4495
László Lipták	84	103	7.1421
Louis R. Bragg	117	159	7.2116
Anna Spagnuolo	150	207	7.9383
Ravindra Khattree	171	207	7.5203
David J. Downing	155	224	7.9347
Sze-Kai Tsui	187	245	7.9876
Louis Jack Nachman	196	265	8.0325
Marc J. Lipman	237	333	7.9221
Eddie Cheng	263	384	8.1200
Hyungju Park	266	395	8.4561
Irwin E. Schochetman	296	422	8.5321
James H. McKay	301	425	8.1576
Charles Ching-an Cheng	319	483	8.3218
Darrell Schmidt	381	507	8.8641
Stephen J. Wright	413	530	8.8550
Stuart S. Wang	386	597	8.2806
Baruch Cahlon	399	644	8.5870
Kevin T. Andrews	431	648	8.8418
Peter Shi	500	802	9.1075
Devadatta M. Kulkarni	499	834	8.8448
Serge Kruk	716	1032	9.3886
J. Barry Turett	434	1056	8.6478
Jerrold Grossman	1026	1566	9.6836
Bo-nan Jiang	1120	2501	9.1857
Meir Shillor	872	2981	8.9189

Appropriate metrics must be devised to assess the joint impact of a group. These metrics can also be useful when hiring new group members: *e.g.* one might be interested in hiring those candidates who most boost the relevance of the group, which differs from the usual criterion that one should hire those candidates with whom group members have already been collaborating.

Definition 6. *The* group citation count *in a certain subject matter is the number of papers published in that subject matter which cite at least one member of the group. This count can be normalized by dividing it by the number of papers published in the subject matter, in order to obtain the* group citation fraction*, which takes values in* $[0, 1]$.

We can also adapt the indices given in Definitions 3, 4 and 5 for groups of authors.

Definition 7. *The* maximum co-cited count of a group of authors $\{i^1, \cdots, i^g\}$ *is* m *if and only if there exists a set of authors* $\{i_1, \cdots, i_m\}$*, where* $\{i_1, \cdots, i_m\} \cap \{i^1, \cdots, i^g\} = \emptyset$ *such that the following conditions hold:*

- *Each author in* $\{i_1, \cdots, i_m\}$ *is co-cited with at least one author in* $\{i^1, \cdots, i^g\}$*;*
- *For every* $1 \leq k \leq m$*, there exists an author* $i(i_k) \in \{i^1, \cdots, i^g\}$ *such that in the weighted co-citation graph the edge* $(i(i_k), i_k)$ *has maximum weight among those edges incident to* i_k*;*
- *No strict superset of* $\{i_1, \cdots, i_m\}$ *verifies the above conditions.*

Definition 8. *The* weighted maximum co-cited count of a group of authors $\{i^1, \cdots, i^g\}$ *is* $\sum_{k=1}^{m} c(i(i_k), i_k)$ *if and only if there exists a set of authors* $\{i_1, \cdots, i_m\}$*, where* $\{i_1, \cdots, i_m\} \cap \{i^1, \cdots, i^g\} = \emptyset$ *such that the following conditions hold:*

- *Each author in* $\{i_1, \cdots, i_m\}$ *is co-cited with at least one author in* $\{i^1, \cdots, i^g\}$*;*
- *For every* $1 \leq k \leq m$*, there exists an author* $i(i_k) \in \{i^1, \cdots, i^g\}$ *such that in the weighted co-citation graph the edge* $(i(i_k), i_k)$ *has maximum weight among those edges incident to* i_k *and this weight is* $c(i(i_k), i_k)$*;*
- *No strict superset of* $\{i_1, \cdots, i_m\}$ *verifies the above conditions.*

Definition 9. *Given a group of authors* $G = \{i^1, \cdots, i^g\}$*, let* $w_1(G), \cdots, w_{n_G}(G)$ *be the weights of the edges connecting members of* G *with non-members of* G *in the weighted co-citation graph. Define* $W(G)$ *as the sum of the previous weights and the relative weights as* $rw_k(G) = w_k(G)/W(G)$*, for* $k = 1$ *to* n_G*. Consider the set of relative weights as a probability distribution over the non-members of* G *co-cited with members of* G*. The Shannon entropy* $H(\{rw_1(G), \cdots, rw_{n_G}(G)\})$ *of that distribution is the co-citation entropy of* G*.*

Indices in Definitions 7, 8 and 9 have the same interpretation as the respective indices in Definitions 3, 4 and 5: the higher their values, the better. Hence, we do not discuss them further.

We will, however, provide some discussion on the group citation count (Definition 6). The following lemma is straighforward but enlightening.

Table 4: Relevance indices of the Oakland Mathematics Department and the Princeton Mathematics Departments as groups

	Maximum co-cited count	Weighted maximum co-cited count	Co-citation entropy
Oakland group	560	1227	12.2070
Princeton group	10207	31764	11.9721

Lemma 1. *Given a group of authors* $G = \{i^1, \cdots, i^g\}$*, where each author* $i^k \in G$ *has received* $c(i^k)$ *citations and* $A(i^k)$ *is the set of articles citing* i^k*, the group citation is maximized if and only if* $A(i^k) \cap A(i^l) = \emptyset$ *for all* $i^k, i^l \in G$ *with* $i^k \neq i^l$*. In this case the group citation count is* $\sum_{k=1}^{g} c(i^k)$*.*

Hence, a group is optimal in the group citation count sense if there is no overlap in the papers citing different group members. When hiring new group members within a certain subject matter, this contradicts the usual idea that one should hire new researchers who are already related via collaboration to the group members: a joint paper by several group members implies subsequent citation overlap, because a paper citing that joint paper will cite several group members. In fact, pushed to the limit, maximization of the group citation count could be seen as discouraging collaboration between group members. Collaboration between group members would only be "rational" if their interaction resulted in a qualitative leap in their joint paper, in such a way that this joint paper would attract more citations than the sum of citations that the contributions of each author to the joint paper would separately attract as independent papers.

Some famous institutions *de facto* follow a pattern of activity which is not very far from the one sketched above. Their model is to hire a limited number of permanent faculty in several areas, who tend to do research in collaboration with a large community of external or visiting co-authors. For example, in the Princeton Institute for Advanced Study, a permanent faculty of no more than twenty-eight academics each year awards fellowships to some 190 visiting members from about one hundred universities and research institutions throughout the world [3].

In Table 4 we show the above group relevance indices computed for the Oakland Mathematics Department and the Princeton Mathematics Department.

5 Concluding remarks

Human evaluation criteria can always be more refined and rich than automated indices. However, when comparing authors or groups, for instance in competitions for work positions or for funding, it is not always possible to analyze one by one the different papers and the production of each candidate (especially if there are many of them). In such cases automated indices may be helpful.

Along this line of thought, we have introduced co-citations and the co-citation graph as new tools to measure the impact of the work by scientific authors. We have argued that the co-citation graph may in fact be more informative than the collaboration graph in gauging the scientific impact: indeed, the co-citation graph gives an idea about how the community perceives and classifies the work by an author, beyond the collaborations that the author has pursued during her/his career.

Co-citations, co-citation graphs and the proposed indices are useful to assess the relevance of individual authors and also the relevance of groups of authors. It has been argued that the relevance of a group of authors is not the sum of the relevances of individuals in the group. Indeed, new indices such as the proposed ones are required in order to measure not only the output of individual authors, but also the output of any research organization, from research groups to entire national or corporate research communities. Open research issues include devising or enhancing the proposed relevance indicators.

Disclaimer and acknowledgments

The authors are with the UNESCO Chair in Data Privacy, but the views expressed in this paper are their own and do not commit UNESCO. The second author is partly supported as an ICREA-Acadèmia researcher by the Government of Catalonia. This work was partly funded by the European Commission under FP7 project "DwB", by the Spanish Government through projects TSI2007-65406-C03-01 "E-AEGIS", TIN2009-11689 "RIPUP" and CONSOLIDER INGENIO 2010 CSD2007-0004 "ARES", and by the Government of Catalonia under grant 2009 SGR 1135.

References

[1] P. Ahlgren, B. Jarneving, and R. Rousseau. Requirements for a cocitation similarity measure, with special reference to pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, 54(6):550–560, 2003.

[2] V. Batagelj and A. Mrvar. Some analyses of Erdős collaboration graph. *Social Networks*, 22(2):173–186, 2000.

[3] Institute for Advanced Study. Mission and history. http://www.ias.edu/about/mission-and-history.html.

[4] J. W. Grossman. The Erdös number project. http://www.oakland.edu/enp.

[5] J. W. Grossman. Patterns of collaboration in mathematical research. *SIAM News*, 35(9), 2002.

[6] J. W. Grossman. Patterns of research in mathematics. *Notices Amer. Math. Soc.*, 52(1):35–41, 2005.

[7] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572, 2005.

[8] L. Leydesdorff. Similarity measures, author cocitation analysis, and information theory. *Journal of the American Society for Information Science and Technology*, 56(7):769–772, 2005.

[9] M. E. J. Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA*, 98(2):404–409 (electronic), 2001.

[10] M. E. J. Newman. A study of scientific collaboration networks i. network construction and fundamental results. *Phys. Rev. E*, 64(016131), 2001.

[11] M. E. J. Newman. A study of scientific collaboration networks ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64(016132), 2001.

[12] M. E. J. Newman. Who is the best connected scientist? A study of scientific coauthorship networks. In *Complex networks*, volume 650 of *Lecture Notes in Phys.*, pages 337–370. Springer, Berlin, 2004.

[13] H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, 1973.

[14] American Mathematical Society. Mathscinet: Collaboration distance. http://www.ams.org/mathscinet/collaborationDistance.html.