J Braz Comput Soc (2011) 17:255–266
DOI 10.1007/s13173-011-0041-7

ORIGINAL PAPER

Ranking in collaboration networks using a group based metric

Vinícius P. Freire · Daniel R. Figueiredo

Received: 23 February 2011 / Accepted: 15 September 2011 / Published online: 12 October 2011

Abstract Collaboration networks are social networks in which relationships represent some kind of professional collaboration. The study of collaboration networks can help identify individuals or groups that are important or influential within a given community. We start this work by characterizing the structural properties of the scientific collaboration network in the area of Computer Science. In particular, we consider the global network (all individuals) and the Brazilian network (individuals affiliated with Brazilian institutions) and establish a direct comparison between them. Our empirical results indicate that despite exhibiting features found in most social networks, these two networks also have some interesting differences. We then present a novel approach to rank individuals within a group in the network (as opposed to ranking all individuals) using solely their relationships. Intuitively, the importance assigned to an individual by our metric is proportional to the intensity of its relationship to the outside of the group. We use the proposed approach and other classical metrics to rank individuals of the Brazilian network and compare the results with the ranking of the Research Fellowship Program of CNPq (an agency of the Brazilian Ministry of Science and Technology). The direct comparison indicates the effectiveness of the proposed approach in identifying influential researchers, in particular when considering top ranked individuals. We then extend the proposed approach to rank small groups of individuals (as opposed to single individuals). We apply this and other classical metrics to rank graduate programs in Computer Science in Brazil and compare the results with the ranking of graduate programs provided by CAPES (an agency of the Brazilian Ministry of Education). Our results indicate that the proposed method can effectively identify influential groups such as well-established graduate programs in Brazil.

Keywords Collaboration networks · Network structure · Node ranking

1 Introduction

Social networks are an important abstraction that has been studied for decades by several areas of knowledge for various reasons [16]. In general, a social network can be represented by a graph (network) $G = (V, E)$, where V denotes the set of individuals under consideration and E the set of relationships that exists among these individuals. A social network can encode one or more of several types of relationship, such as friendship, kinship, sexual contact or professional collaboration. Intuitively, relationships have different intensities that reflect the strength of the social tie. For example, consider the number of phone calls placed in a year between two individuals as a metric to reflect the intensity of this social tie. The intensity of a relationship is usually represented by a function $w(e)$, $e \in E$, that associates a weight to each edge of the network.

A scientific collaboration network is a social network where vertices correspond to authors of scientific papers and edges between authors exist if they have published at least one paper together. A simple metric to measure relationship

A preliminary version of this paper appeared at the Brazilian Symposium on Collaborative Systems (SBSC) 2010.

V.P. Freire · D.R. Figueiredo (✉)
Alberto Luiz Coimbra Institute for Graduate School and Research in Engineering (COPPE), Systems Engineering and Computer Science Program (PESC), Federal University of Rio de Janeiro (UFRJ), Caixa Postal 68511, Rio de Janeiro 21941-972, Brazil
e-mail: daniel@land.ufrj.br

V.P. Freire
e-mail: vini@land.ufrj.br

intensity in this network is the number of papers two individuals published together. Several topological properties can be characterized in a collaboration network which can then be used for different purposes. For example, a common task is to rank researchers according to their importance or groups of individuals according to their interest [13, 15, 18]. However, these works are based on metrics that consider vertices and their incident edges in isolation and provide a full ranking of all vertices.

Motivated by the problem of ranking researches, we propose a new approach to measure the importance of individuals and groups of individuals. The key idea is to capture importance not as a whole (within the entire network), but within a relative small group of individuals. Moreover, the importance of an individual within the group is proportional to the intensity of its relationship to individuals outside the group. Intuitively, this captures the importance of an individual with respect to exchanging information with the outside of the group, serving as a bridge for the group. In this regard, we make three distinct contributions:

- We present a study of various topological properties of two collaboration networks formed by authors of scientific papers in Computer Science: global collaboration network (all authors) and Brazilian collaboration network (only authors affiliated with Brazilian institutions). We characterize the metric used to measure relationship intensity in these two networks and establish a direct comparison between them, illustrating their similarities and differences.
- We apply the proposed approach to rank individuals in collaboration networks by considering the set of Brazilian authors as the target group. We evaluate the effectiveness of our approach by comparing this ranking with other previously proposed ranking metrics for collaboration networks. In particular, we compare it with the ranking of the Research Fellowship Program of CNPq (National Council for Scientific and Technological Development) which grants fellowships to Brazilian researchers. Our results indicate that the proposed metric is effective in identifying the influential researchers in Brazil, specially when the considering top ranked individuals.
- We extend the proposed metric to rank groups of individuals. We consider groups that correspond to faculty associated with graduate programs in Computer Science in Brazil and rank these groups according to various metrics. We compare these rankings with the highly visible ranking of Brazilian graduate programs provided by CAPES (an agency of the Ministry of Education in Brazil). Our results show that the proposed metric can accurately identify established graduate programs in Brazil, indicating its usefulness in identifying influential groups of individuals.

The remainder of this paper is organized as follows. In Sect. 2 we present the related work and some discussions. Section 3 presents details of the collaboration network studied and the characterization of its properties, including metrics for measuring relationship intensity. Section 4 introduces our approach for ranking individuals within a group and presents several rankings of Brazilian researchers according to different metrics and an empirical evaluation of their effectiveness. In Sect. 5 we extend our approach to rank groups of individuals and establish a comparison with other metrics and evaluate their performance. Finally, Sect. 6 concludes the paper.

2 Related work

Scientific collaboration networks have been studied for decades both through scientific perspective as well as popular culture [2, 16]. For example, the *Erdős number* of an individual represents the distance to the famous mathematician Paul Erdős in the scientific collaboration network [5]. This has been widely disseminated with some researchers proudly reporting their Erdős number.[1]

Scientific studies concerning collaboration networks have focused on characterization and comparison of structure in different areas of knowledge [14]; characterization and comparison of geographically distinct groups of researchers [11]; evolution of the collaboration network over time [1, 6, 11]; mathematical models for collaborations network and prediction of future collaborations [6, 9].

Another problem studied in the context of collaboration network is ranking of individuals according to their importance. The problem is to rank vertices using solely structural information and possibly edge weights representing relationship intensity. Newman introduced a simple metric to measure the intensity of relationships in scientific network, which works as follows [15]. Every paper written by a set of k authors adds $\frac{1}{k-1}$ to the weight of every edge among these k authors. More formally, let P denote the set of publications under consideration (i.e., some publication database). Let A_p denote the set of authors of publication $p \in P$. Let V denote the set of authors that appear in at least one publication, thus, $V = \cup_{p \in P} A_p$. We can now define the edge weight function, as follows:

$$w(u, v) = \sum_{p \in P} \frac{\mathbf{1}(u, v \in A_p)}{|A_p| - 1} \quad \text{for all } u, v \in V, \tag{1}$$

where $\mathbf{1}()$ is the indicator function and $|X|$ denotes the size of set X. Finally, let E denote the edge set of the network and be defined by all unordered pairs (u, v), $u, v \in V$, that have positive weight, $w(u, v) > 0$.

[1] The Erdős number of the first and second author of this paper is 5 and 4, respectively.

Note that from the point of view of an author, every paper coauthored will add a weight sum of 1 divided among all other coauthors of that paper. For example, if a paper is coauthored by authors A, B and C, then the weight of the edge between A and B will increase by 0.5 as well as the weight of the edge between A and C. Note that this metric is more robust than the simple metric that counts the number of papers coauthored by two individuals. In this simple metric, the number of authors of each paper is not considered, and a single paper can bring a lot of weight to the network. On the other hand, in Newman's metric each author contributes a normalized amount of effort to a paper (unity), independent of the number of authors of the paper.

Using this metric for edge weights, Newman determines the vertex weight as the sum of the weights of edges incident to the vertex [15]. In particular, for all $v \in V$ we define

$$w(v) = \sum_{e \in E, e=(u,v)} w(u, v), \tag{2}$$

where E is the set of edges of the network. Note that this simply corresponds to the number of papers that the vertex $v \in V$ has in collaboration with any other person. Using $w(v)$, Newman ranks vertices in decreasing order, establishing the most "influential" scientists across different communities [15]. Unfortunately, Newman did not qualitatively assess the performance of his approach.

The main difference between the approach proposed here and Newman's approach is that we will not consider all edges incident to a vertex when determining the vertex weight. In particular, only edges of a certain kind, which relates to the notion of group, will be considered for determining the vertex weight (this will soon become clear). Moreover, we will only rank individuals within a group and the ranking will be relative to the group and not absolute in the network.

3 Characterizing the collaboration network

The scientific collaboration network used in this article was built using the *DBLP* (*Digital Bibliography & Library Project*) database obtained in June 2009. *DBLP*[2] is a publicly available database centrally managed with bibliographic information of key journals and conferences in the Computer Science, with over 1.3 million publications and 750 thousand authors (in June 2009). It is a world reference and widely used by the academic community to search for bibliographic information of Computer Science publications [8]. However, *DBLP* is not a comprehensive database of papers published in Computer Science. For example, it has limited information concerning Brazilian conferences and journals and does not provide uniform coverage across subareas of Computer Science (e.g., theory is usually underrepresented). An advantage of *DBLP* is its central management, which yields fairly accurate information and handles multiple name issues (authors that appear in publications with different names are identified and considered the same person) and its good coverage over some important subareas of Computer Science [7, 11, 12].

In the scientific collaboration network built, each vertex corresponds to an author registered in *DBLP* and there is an edge between two authors if they are coauthors in at least one publication registered in *DBLP*. The weights of the edges are calculated according Newman's metric, as described in Sect. 2.

3.1 The Brazilian network

The Brazilian network is an induced subgraph of the global collaboration network where every vertex is affiliated with a Brazilian institution. Thus, only these researches and collaborations among them form the Brazilian network. Unfortunately, *DBLP* does not classify authors based on their nationality or affiliation, thus we developed a method to obtain the set of authors that are affiliated with a Brazilian institution. We started by using the URL of the homepage of authors available in *DBLP*. If this URL ended with ".br/" then we included the author in the Brazilian network. Unfortunately, only about 200 authors were identified using this criterion.

In order to obtain more authors for the Brazilian network, we considered all faculty of all graduate programs in Computer Science in Brazilian universities as listed by CAPES[3] and all researchers that receive a research fellowship from CNPq[4] (both are publicly available). Individuals in this list that were identified in *DBLP* were placed in the Brazilian network. Automatic name variations were also considered to increase the match in *DBLP*, since Brazilian names tend to appear with several variations in bibliographic databases. This increased the size of the Brazilian network to about 1600.

Finally, we inspected the neighboring individuals in the global network of authors already in the Brazilian network to verify if they were also affiliated with a Brazilian institution. This verification was done manually by visiting the person's personal or institutional website. If the person was found to work in Brazil, he or she was placed in the Brazilian network. Finally, the list of Brazilian authors came to 2,729 researchers.

[2] Available at http://www.informatik.uni-trier.de/~ley/db/.

[3] Available at http://www.capes.gov.br/.

[4] Available at http://www.cnpq.br.

Table 1 Summary of structural properties of the two collaboration networks

Metric	Global	Brazilian
Number of authors	722,392	2,729
Number of edges	2,272,540	6,953
Edge density	8.7×10^{-6}	1.8×10^{-3}
Number of publications	1,230,213	13,314
Number of pub/author	1.7	4.9
Degree range	[0, 643]	[0, 101]
Mean degree	6.3	5.1
Size of GCC	576,309	2,338
Relative size of GCC	79.8%	85.7%
Size of 2nd largest CC	42	13
Number of CC	77,493	297
Mean size of CC	9.3	9.2
Clustering coeff.	0.59	0.48
Mean distance	6.3	5.6
Diameter	23	15
Mean edge weight	0.63	1.09
Edge weight range	[0.0088, 267.8]	[0.03, 86.6]
Mean vertex weight	3.9	5.6
Vertex weight range	[0, 529]	[0, 123]
Mean publ. age	8.26	5.46
Oldest publ.	73	38

3.2 Structural analysis

We now present a detailed characterization of various structural properties of the scientific collaboration network, comparing the global network with the Brazilian network. We start by noting the very different sizes of these networks, with the global network having 722,392 authors and 2,272,540 edges (collaborations) while the Brazilian network has 2,729 authors and 6,953 edges. Note that the Brazilian network is a rather small induced subgraph of the global network. However, the Brazilian network is relatively much more dense than the global network (i.e., the fraction of edges it has over all possible edges). Another aspect is that the average number of publications per author is also much higher within the Brazilian set of authors. However, the Brazilian network has a smaller average degree (i.e., smaller average number of collaborators). A summary with the average values of all metrics considered are shown in Table 1.

These observations about the Brazilian network can be misleading. In one hand, it seems that Brazilians publish more with fewer collaborators in average, which is certainly what the data analysis indicates. However, recall that the Brazilian list of researchers was obtained using the faculty of graduate programs and fellowship recipients, leading to a more selective group (i.e., no post-docs or students). Moreover, *DBLP* has limited coverage of Brazilian venues, where large number of Brazilian researchers publish. Thus, in some sense the collected list of Brazilians has a bias toward individuals that tend to publish more and in international venues, as they appear in *DBLP*.

Degree

Let d_u denote the degree of vertex $u \in V$ in the network, which corresponds to the number of collaborators of individual u. The empirical complementary cumulative distribution function (CCDF) of the degree [2, 16] for both networks are shown in Fig. 1. Note that the x-axis represents the degree, i.e. the number of coauthors, and the y-axis is the fraction of vertices with degree greater than or equal to x.

Figure 1 indicates that the degree distribution for both networks exhibits a heavy tail, with a wide range of values that occur far from the mean degree (see range and mean in Table 1). Heavy tail is a common feature in social and collaboration networks, and was observed in prior studies [6, 11, 14, 17, 18]. Note that there are a small number of authors with many collaborators and a lot of authors with few collaborators. For example, in the global network we observe that 15% of vertices have degree greater than 10 and 80% of vertices have degree less than 8.

The degree distribution also reveals some unexpected feature: 6% of authors in the global network do not have any collaborators (registered in *DBLP*), corresponding to more than 43,000 authors. In the Brazilian network, 9% do not have any collaborators (in the induced subgraph), corresponding to 243 authors. Thus, these vertices have no incident edges in the collaboration network.[5]

Clustering coefficient

The clustering coefficient c_u of vertex u measures the connectivity between the neighbors of u [2, 16] and captures the relative number of triangles in the network. In particular, c_u is obtained as follows:

$$c_u = \frac{E_u}{d_u \times (d_u - 1)/2} \tag{3}$$

where E_u is the number of edges among the neighbors of vertex u and $\binom{d_u}{2}$ is the maximum number of edges among them. Also, if $c_u = 0$ if $d_u \leq 1$.

The clustering coefficient of a network is just the arithmetic mean of the coefficients of all its vertices. In the global collaboration network, the clustering coefficient is 0.59, while for the Brazilian network this value is 0.48, which is 19% smaller. In either case, the chances that two

[5] The point $x = 0$ does not appear in the figure due to the log–log scale of the plot.

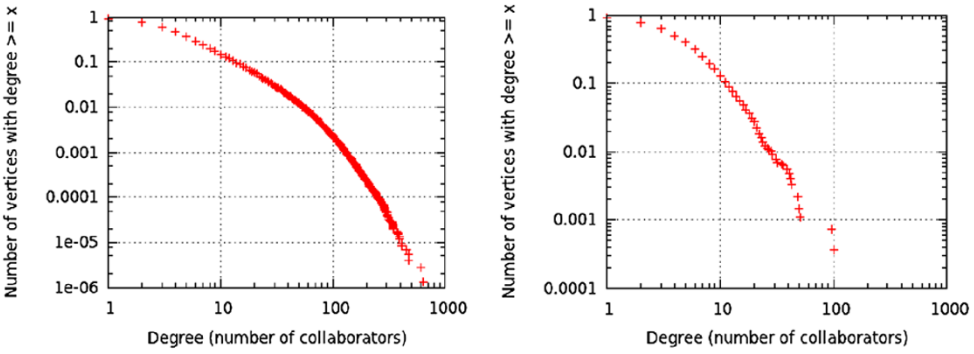


Fig. 1 Degree distribution (CCDF) of collaboration networks

(a) Global network. (b) Brazilian network.

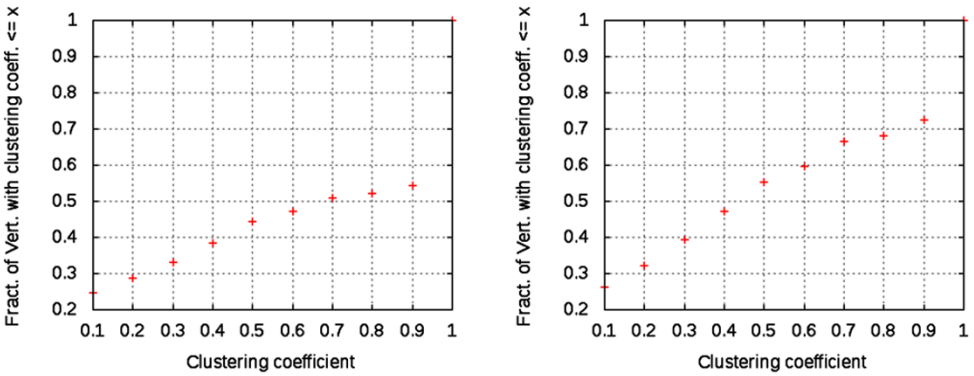


Fig. 2 Clustering coefficient distribution (CDF) of collaboration networks

(a) Global network. (b) Brazilian network.

authors that have a common collaborator will also collaborate with each other is relatively high. This characteristic is commonly found in many social networks, including various collaboration networks [16]. Finally, the clustering coefficient distribution is shown in Fig. 2, where we observe for the global network that 46% of the vertices have a clustering coefficient greater than 0.9 and only 25% of vertices has a clustering coefficient lower than 0.1. The Brazilian network has lower clustering, specially when considering the fraction of vertices with very high clustering coefficients. In any case, we conclude that the vast majority of vertices have many triangles around them.

Distance

The distance between two vertices in the collaboration network is simply the length in hops of the shortest path between the vertices. The average distance of the network is given by the arithmetic mean of the distance between all pairs of vertices that have a distance defined (i.e., belong to the same connected component) [16]. The global network has an average distance of 6.3 while the Brazilian network of 5.6. Note that these are rather very small values, given the size of the network, supporting the general notion that social networks form a "small world" [2, 16].

The degree distribution is shown in Fig. 3, where the average distance can be clearly identified. We also note that the vast majority of vertices have short distances and that there are no heavy tails. The largest distance in the network, known as the diameter, is 23 and 15, in the global and Brazilian networks, respectively. Note again that the diameter is rather small compared to the network size, a feature also found in other social networks [16].

Edge weight

We report on the relationship intensity (edge weights) distribution as computed using Newman's metric, described in Sect. 2. Figure 4 shows the distribution for both collaboration networks. We observe that both cases exhibit a heavy tail, showing a wide range of values and a very small average (see details in Table 1). By inspecting *DBLP* we find out that the heaviest edge is among two authors that have published 336 papers together, out of which 224 had only the two as coauthors. The lightest edge is between two people that have coauthored just a single paper together, but with a total of 114 authors (edge weight of 1/113). Within the Brazilian network, the heaviest edge is between two authors that have written 100 papers together, out of which 80 had only the two as coauthors. The lightest edge comes from a paper with 29 authors.

The plots in Fig. 4 show two distinct extremes: authors that collaborate very infrequently and with several other

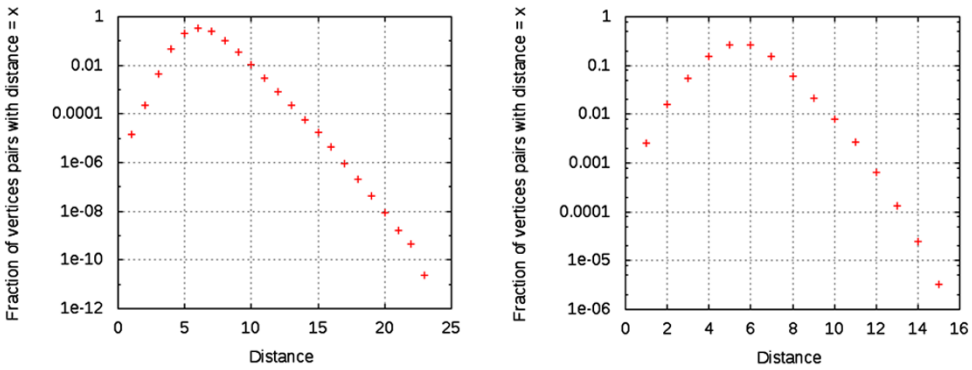


Fig. 3 Distance distribution (PDF) of collaboration networks

(a) Global network. (b) Brazilian network.

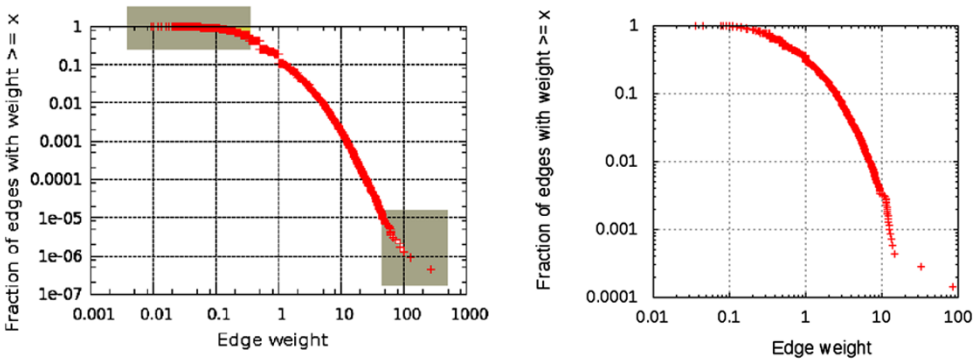


Fig. 4 Edge weight distribution (CCDF) of collaboration networks

(a) Global network. (b) Brazilian network.

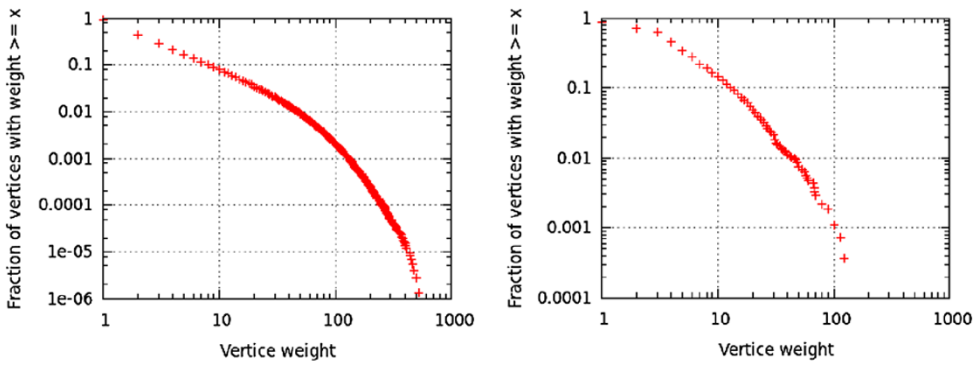


Fig. 5 Vertex weight distribution (CCDF) of collaboration networks

(a) Global network. (b) Brazilian network.

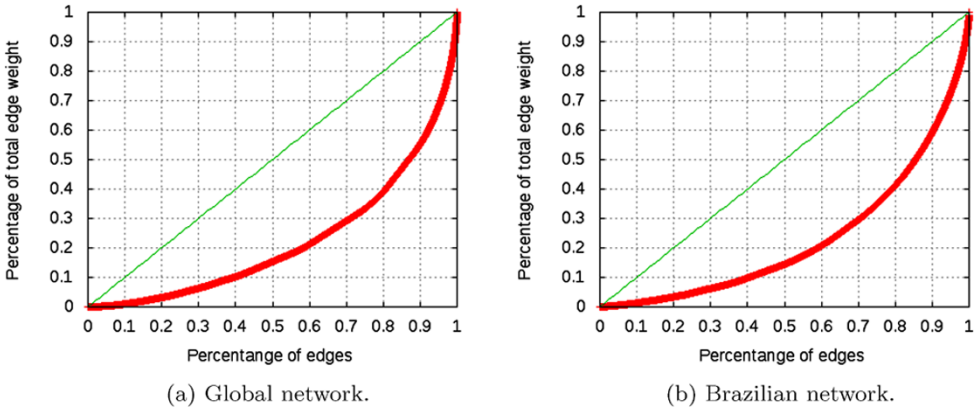
coauthors; authors that collaborate very frequently with few other coauthors. Such behavior is present in both global and Brazilian networks and seems to be a fundamental characteristic of scientific collaboration. Finally, the discontinuities in values such as $x = 1.0, 0.5, 0.333, 0.25, 0.2, 0.166$ occurs because most papers have 2, 3, 4, 5, 6 or 7 authors, significantly increasing the frequency of edges with these values.

Vertex weight

Recall that the vertex weight corresponds to the number of publications that an individual has with at least one more coauthor. Figure 5 shows the distribution of vertex weights. Again, note that this characteristic also exhibits a heavy tail, with average values much smaller than their range (details in Table 1). Note that only 10% of the authors have collaborated in more than eight articles and within the Brazilian network this number goes up to 19%. Thus, the vast majority of authors do not have many publications in collaboration, while very few authors collaborate significantly with others.

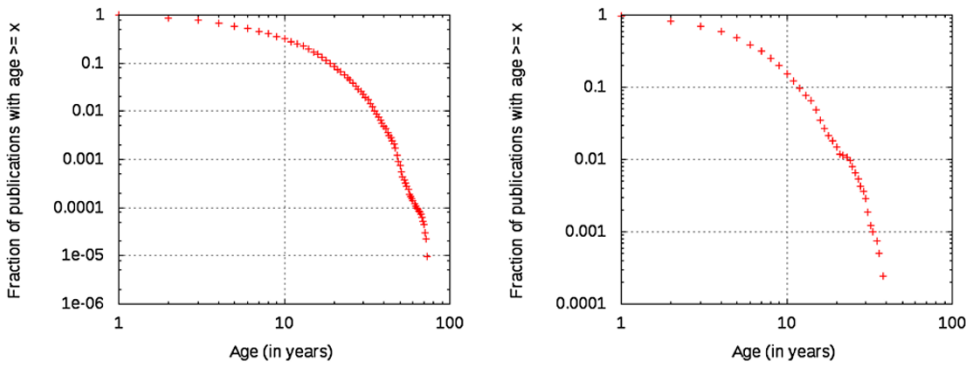
Gini coefficient for edge weight

As with several other social phenomena, we have shown that intensity of relationships are also not uniformly distributed across a population (in this case, across edges). In fact, most



(a) Global network. (b) Brazilian network.

Fig. 6 Lorenz curves for the distribution of edge weights in collaboration networks



(a) Global network. (b) Brazilian network.

Fig. 7 Publication age distribution (CCDF) of collaboration networks

relationships are very weak while very few are extremely strong, as shown in Fig. 4. In order to characterize this inequality, we will use the Lorenz curve and the Gini coefficient [4]. Recall that the Gini coefficient is a number between 0 and 1 that measures the inequality of a distribution with higher values being more unequal (0 represents the uniform distribution).

Figure 6 shows the Lorenz curves for the edge weight distribution. The x-axis represents the percentage of edges being considered, sorted from lightest to heaviest. The y-axis corresponds to percentage of accumulated edge weight. Considering the global network, we note that 80% of the edges with less weight accumulate 39% of the total edge weight. Thus, 20% of the edges are responsible for 61% of all edge weights, showing the inequality of this distribution. The Brazilian network exhibits a similar trend, indicating that this social phenomenon is probably inherent in collaboration networks. Finally, the Gini coefficient for the global and Brazilian networks are 0.55 and 0.54, respectively, indicating that both weight distributions are very unequal.

Publication age

The age of a publication refers to the number of years since it was published. Since our database was collected in July 2009, all papers published in 2009 have age 0. Thus, a paper published in year n has age $2009 - n$. Using this metric, we compute the empirical distribution of the age of publications in *DBLP*, as shown in Fig. 7. We note that most publications are quite young in both networks, particularly in the Brazilian network. For example, 68% are less than 10 years old, going up to 85% when considering the Brazilian network. This is an indication of the growth in publishing in Computer Science, possibly coupled with an increase in the coverage of *DBLP*. Clearly, the Brazilian network is younger than the global network, reflecting Brazil's recent (last 10 years) strong academic growth [7].

Collaborators versus publications

We consider the correlation between the number of collaborators and number of publications of individual authors. To investigate this issue, we present a scatter plot of the degree versus the number of publications where a point in the plot corresponds to some number of authors. Figure 8 shows this scatter plot for the global network (note that the axis are in log scale and that binning was also done in log scale). Note that there is a large concentration of authors near the main diagonal (light colors) indicating a strong correlation between the two properties: the larger the number of collaborators the larger the number of publications. Interestingly, there are several exceptions, both with high number

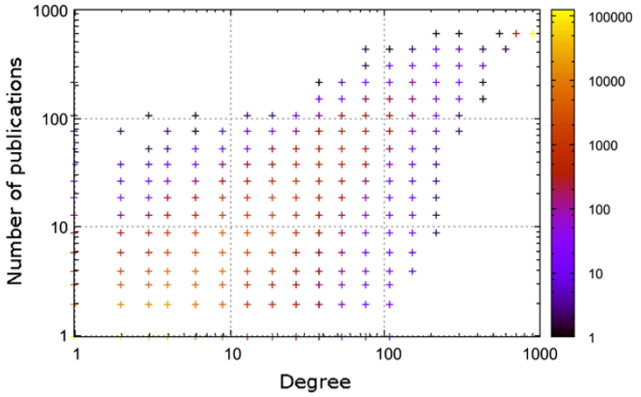


Fig. 8 Number of collaborators and publications of individuals (*log scale*)

of collaborators and low number of publications, and also low number of collaborators and high number of publications. For example, there are two authors that have a single collaborator but over 200 publications; while there are forty authors that have a single publication, but with over 100 collaborators.

4 Ranking of individuals in a group

In this section we introduce our metric to rank individuals in collaboration networks. Our approach is based on two key ideas: (i) consider a relative small set (group) of individuals and rank them within this set; (ii) the importance of an individual to a group is proportional to the intensity of its relationships with individuals *outside* the group. Intuitively, important individuals in a group play the role of "bridges" between the group and the outside world. Through these individuals, ideas and knowledge flows to and from the group, which is an important aspect when considering scientific collaboration networks. On the other hand, individuals that have no strong relationships with individuals outside the group are likely to have smaller importance within the group.

The metric is based on cuts of graphs and weight of vertices in the cut. Let $G = (V, E)$ represent a graph corresponding the collaboration network and let X be a subset of the vertices, $X \subset V$. The cut C induced by the set X is given by the set of edges that have one endpoint in X and the other in $V - X$. Thus,

$$C = \big\{(u, v) \mid (u, v) \in E, u \in X, v \in V - X\big\}. \tag{4}$$

The weight of cut C is given by the sum of the edge weights present in the cut. Thus,

$$w(C) = \sum_{e \in C} w(e)\,, \tag{5}$$

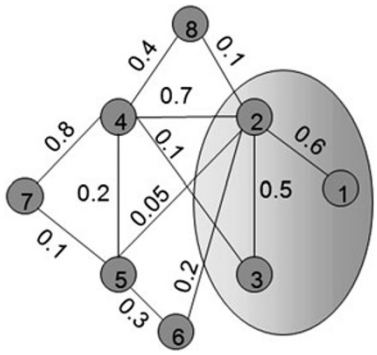


Fig. 9 Example illustrating the vertex weight in the cut

where $w(e)$ is the intensity of a relationship (edge weight), as defined in (1).

We now redefine the weight of a vertex to consider only its relationships with individuals outside the group. Let $q(v)$ denote the contribution of vertex v to the weight of the cut C, where $v \in X$. Thus,

$$q(v) = \sum_{e \in C, e=(u,v)} w(e), \quad \text{where } v \in X. \tag{6}$$

Note that if $v \in X$ has no relationships with individuals outside of X, then $q(v) = 0$. Another observation is that the sum of all $q(v)$ yields the cut weight. Thus,

$$w(C) = \sum_{v \in X} q(v). \tag{7}$$

The example illustrated in Fig. 9 helps clarify our approach. The example network has eight vertices with the set $X = \{1, 2, 3\}$, as shown in the figure. The cut weight is $w(C) = 1.15$ and the contribution of vertices in X to this cut weight is given by $q(1) = 0$, $q(2) = 1.05$, $q(3) = 0.1$.

Finally, we will use the contribution of each vertex to the weight of the cut as the ranking metric. Thus, vertices in X will be ranked according to $q(v)$, in decreasing order. The vertex $v \in X$ with the largest $q(v)$ will be ranked as the most important vertex in the group, and so on. In the previous example, vertex 2 would be ranked first, with vertex 3 coming second, followed by vertex 1.

Intuitively, our approach is adequate when X is a community or a set of individuals that have relationships among them. Our approach is not suitable if X is any set of individuals, like a random set. Moreover, our approach is more adequate when $|X| \ll |V|$, that is, when the number of individuals in the set is much smaller than the network size. In the following, we will work under both assumptions and evaluation of the approach when these conditions do not hold is left for future work.

4.1 Evaluation of different ranking metrics

In this section we will rank Computer Science researchers affiliated with Brazilian institutions (i.e., all vertices in the Brazilian network) using four different metrics, including the approach proposed above. Our goal is to compare and evaluate the quality of the ranking produced by the different metrics.

To assess the quality of the rankings we will use information from the Research Fellowship Program of CNPq,[6] which grants fellowships to researchers affiliated with Brazilian institutions. Fellowships belong to one of two categories, 1 and 2, with category 1 subdivided into four levels, A, B, C and D. The different categories are used to reflect seniority, productivity and impact of researchers and is also related to the fellowship monetary value. Category 2 serves mostly young researchers, while category 1 requires at least eight years since obtaining the doctoral degree. Category 1A is the most prestigious and is reserved for researchers that have shown continued excellence in scientific production and training of human resources and are members of consolidated research groups. The list of recipients of fellowships is publicly available and maintained by CNPq.[7] Finally, all fellowship recipients in the area of Computer Science were identified in the *DBLP* database. Note that we used the list of fellowship recipients for the year of 2009, which changes yearly.

The categories of the research fellowships granted by CNPq provide a natural ranking of Brazilian researchers. In fact, such fellowships are quite prestigious in Brazil and highly publicized and perceived by academia as an important criterion to assess researchers' merits. Moreover, researchers in category 1A are perceived as the well-established and senior researchers in their area.

We evaluate the different ranking metrics by assessing the "quality" of the ranked list produced by the metric In particular, we use precision, recall, and F-measure having as baseline researchers with different fellowships from CNPq. Our goal is to evaluate the performance of the metrics in identifying this set of researchers. We remind the reader that precision is the fraction of the set of objects retrieved that are relevant; recall is the fraction of relevant objects that were retrieved; F-measure is the harmonic mean between precision and recall [10].

We consider four different metrics to rank individuals: number of publications; number of collaborators (degree); vertex weight $w(v)$, as defined in (2); and vertex cut weight $q(v)$, when using the set of Brazilian researchers as the group of interest. Recall that data for this analysis was obtained using *DBLP*, as described in Sect. 3. We rank researchers in decreasing order using each of these metrics and consider the top L researchers in the ranking. Using this set of L researchers we determine the number of fellowship recipients of a particular kind that appear in this list. This number is then used to compute the precision, recall and F-measure of each ranking metric.

Table 2 Precision and recall of various metrics in identifying 1A fellowship recipients under different list sizes

L	No. pub.		$w(\cdot)$		$q(\cdot)$		Degree	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
20	0.15	0.14	0.15	0.14	**0.30**	**0.29**	0.20	0.19
25	0.16	0.19	0.12	0.14	**0.28**	**0.33**	0.16	0.19
30	0.13	0.19	0.13	0.19	**0.23**	**0.33**	0.13	0.19
35	0.14	0.24	0.14	0.24	**0.20**	**0.33**	0.11	0.19
50	**0.16**	**0.38**	**0.16**	**0.38**	**0.16**	**0.38**	0.08	0.19

We start by considering researchers that have been granted 1A fellowships as the target group and lists sizes varying from 20 to 50, which corresponds to 0.7% and 1.8% of the Brazilian network. Table 2 presents the precision and recall for all ranking metrics. Values in bold correspond to the highest values of precision and recall for each list size. Note that vertex cut weight $q(\cdot)$, outperforms all other ranking metrics in both precision and recall. When the list size is 50, two other ranking metrics (number of publications and vertex weight $w(\cdot)$) exhibit the same performance as the proposed approach. We also note the relatively high recall (0.33) of the proposed metric when the list size is 25 or larger.

An interesting observation concerning Table 2 is that vertex weight and number of publications have the same value for precision and recall for almost all list sizes. This is due to the similarity between these two metrics since vertex weight corresponds to the number of publications with at least one collaborator (coauthor). We also note that the degree (i.e., number of collaborators) does not identify additional 1A researchers as the list size increases from 20 to 50 (no change in recall), and is therefore a not very robust metric.

We now consider the effectiveness of the metrics in identifying different sets of fellowship recipients by considering different target groups. Table 3 shows the F-measure for all increasing sets of fellowship recipients (different target groups) and various list sizes, with values in bold indicating the highest F-measure for the row. Note that 1(A, B, C, D)2 includes all recipients of fellowships. We observe that vertex cut weight $q(\cdot)$ outperforms other metrics in various scenarios, in particular, when considering list sizes of 20 and 30. Curiously, the metric is outperformed when considering the set 1(A, B, C, D) of fellowship recipients. Finally, when the list size grows to 50, the metric is outperformed by number of publications. This indicates that the proposed metric is more effective in identifying influential individuals when considering the top rankings yielded by the metric.

[6] In Portuguese, Programa de Produtividade em Pesquisa (PQ).

[7] Available at http://www.cnpq.br.

Table 3 F-measure of various metrics in identifying different sets of fellowship recipients under different list sizes

L	Category	pub.	$w(\cdot)$	$q(\cdot)$	deg
20	1A	0.15	0.15	**0.29**	0.19
20	1(A, B)	0.16	0.16	**0.29**	0.23
20	1(A, B, C)	0.24	0.24	**0.26**	0.22
20	1(A, B, C, D)	0.18	0.18	0.20	**0.21**
20	1(A, B, C, D)2	0.09	0.09	**0.10**	0.09
30	1A	0.16	0.16	**0.28**	0.16
30	1(A, B)	0.25	0.25	**0.28**	0.25
30	1(A, B, C)	**0.29**	**0.29**	**0.29**	**0.29**
30	1(A, B, C, D)	0.24	0.25	0.25	**0.30**
30	1(A, B, C, D)2	0.13	0.13	**0.14**	0.13
50	1A	**0.22**	**0.22**	**0.22**	0.11
50	1(A, B)	**0.33**	**0.33**	0.28	0.24
50	1(A, B, C)	**0.39**	**0.39**	0.33	0.33
50	1(A, B, C, D)	**0.36**	0.35	0.35	**0.36**
50	1(A, B, C, D)2	**0.21**	**0.21**	**0.21**	0.20

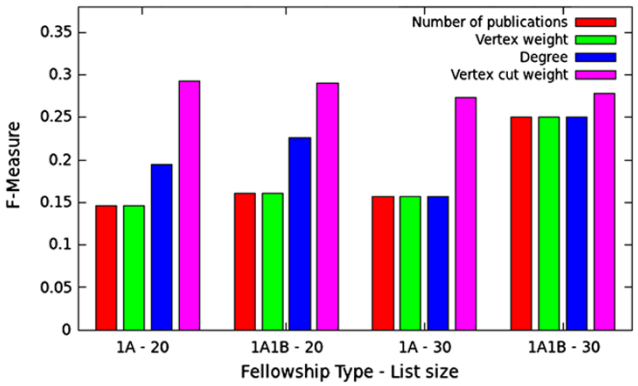


Fig. 10 F-Measure of ranking of four different metrics when identifying 1A and 1B fellowship recipients

Finally, Fig. 10 shows a direct comparison of the F-measure for two baselines (1A and 1(A, B)) and two different list sizes ($L = 20, 30$). The vertex cut weight $q(\cdot)$ outperforms all other metrics and in some cases exhibits an F-measure that is almost twice the value of others. Interestingly, when considering a list size $L = 30$, all other metrics exhibit exactly the same performance, while for list size $L = 20$ degree (which indicates the number of collaborators) has performance superior than the other two metrics, but still inferior to vertex cut weight. A more detailed analysis of the various scenarios can be found in [3].

5 Ranking of groups

In this section we extend the proposed approach to rank groups of individuals. In particular, we are interested in comparing different groups of individuals in scientific collaboration networks and therefore, developing a network-based metric that can be used to rank groups. Using the same intuition as before, we will equate the importance of a group of individuals with the intensity of their relationships with individuals outside the group. Again, the intuition is that groups that have strong ties with the outside are more likely to disseminate and absorb information and ideas and therefore are likely to be more influential, specially when considering groups of researchers in academia.

As before, let X be a subset of the vertices of the scientific collaboration network, $X \subset V$, and let C be the cut induced by X. Recall that the weight of the cut is given by $w(C)$, as given by (5). In order to establish a direct comparison between groups of different sizes, we normalize the group weight by the number of individuals in the group. Thus, we have

$$\bar{w}(X) = \frac{w(C)}{|X|}, \tag{8}$$

where C is the cut induced by the set of individuals in X. We refer to $\bar{w}(X)$ as the average group weight.

Intuitively, the group weight will be an adequate metric when members of group X have some social binding (i.e., collaborations), such as a research group, a project team, or a graduate program. The metric does not seem adequate if there are few (or none) relationships between the members of X, for example, if X is chosen randomly. Finally, we will establish that groups with larger $\bar{w}(X)$ are more influential.

We consider two other metrics to rank groups: (i) number of publications of the group and (ii) number of collaborators (i.e., coauthors) of the group. The number of publications is the number of entries (i.e., publications) that appear in *DBLP* that has at least one coauthor that is a member of the group. Note that a publication may be credited to more than one group, if coauthors belong to different groups. The number of collaborators is the number of individuals outside the group that have coauthored at least one paper (that appears in *DBLP*) with a member of the group. Thus, the number of collaborators represents the "frontier" between the group and the outside. Note that number of collaborators counts edges in the cut C, while $\bar{w}(X)$ considers the weight of these edges. Finally, to allow a direct comparison between groups, we will also normalize these metrics by the group size.

5.1 Evaluation of different ranking metrics

In order to assess the different metrics, we will use as groups faculty members of Computer Science graduate programs in Brazilian universities. The list of faculty members of each graduate program is publicly available and is provided by

CAPES,[8] an agency of the Brazilian Ministry of Education. Among other duties, CAPES continuously assesses the quality of all graduate programs in Brazil. The evaluation occurs every three years and each program receives a score in the range from 3 to 7, where 3 indicates a new or young graduate program and, 6 and 7 indicate international excellence, being reserved for a small percentage of the graduate programs [7]. CAPES's triennial evaluation is publicly available and is highly publicized in Brazil, being used as a criterion by students to select graduate programs and by funding agencies to grant financial support to the programs (including CAPES itself).

CAPES's assessment of the different graduate programs in Computer Science will be used as a baseline to compare the different metrics to rank groups of individuals. We considered 21 different graduate programs spanning all five different scores given by CAPES in the triennial evaluation of 2004/2006 (the latest available at the time of this study). Note that in this evaluation only two and three graduate programs in Computer Science in Brazil had the scores of 7 and 6, respectively. Moreover, the faculty members of each graduate program were also obtained from CAPES and were associated with this same triennial evaluation (thus, it reflects the faculty members of a graduate program at the end of 2006). These faculty members were manually searched for in *DBLP* in order to identify them in the scientific collaboration network. Thus, X_d is the group formed by all faculty members of graduate program d. Note that not all faculty members of a graduate program appear in *DBLP*, since they may not have a publication listed in the database. In this case, they were treated as isolated nodes (no edges) in the scientific collaboration network. However, there were very few such cases and these did not impact the overall result.

Table 4 presents the values of the different metrics for each graduate program studied. The first column identifies the score given by CAPES to the graduate program (first number) as well as a letter to identify the program. The other columns report the total and average values for each metric. We observe that the top programs (ranked 6 and 7 by CAPES) exhibit high values in all average metrics, including average group weight. In particular, we can establish a threshold for each metric such that all and only programs ranked 6 and 7 by CAPES are above this threshold (19 for average number of publications, 17.5 for average number of collaborators, and 19 for average group weight).

Unfortunately, the picture is not as clear for the remainder of the graduate programs. For example, consider the programs ranked 5 by CAPES. They do not form a consistent group in any of the average metrics considered. In particular, for each average metric there is at least one program ranked 3 by CAPES that has a value at least as large as a program ranked 5 (e.g., for $\bar{w}(X)$ 3C is higher than 5B and 5C). However, when considering the absolute values for the metric, all programs ranked 5 have higher values than any program ranked 3. But in this case, we have programs ranked 4 with higher absolute values than programs ranked 5, for all metrics (e.g., 4D is higher than 5B in all metrics in their absolute value). Thus, a single metric cannot reflect the overall ranking established by CAPES across all graduate programs. In particular, it seems a mixture of metrics may be needed to produce an overall ranking that is more in accordance with CAPES. We leave this issue for future investigation.

Figure 11 shows a direct comparison between the average number of publications and average group weight. Each point in the plot corresponds to a graduate program and its shape corresponds to the ranking according to CAPES. All 21 graduate programs studied are shown in the plot. We observe that programs ranked 6 and 7 are clearly isolated from the remainder under either metrics. However, programs ranked 5, 4 and 3 have some mixing in both metrics. In any case, the average group weight is effective in identifying the most influential groups, namely graduate programs ranked 6 and 7. A more detailed analysis and other aspects of the groups can be found in [3].

Table 4 Performance of different metrics for ranking groups of individuals

P	$\|X\|$	#pu	$\overline{\text{pu}}$	#co	$\overline{\text{co}}$	$w(X)$	$\bar{w}(X)$
7A	40	900	22.5	826	20.7	881	22.0
7B	24	855	35.6	454	18.9	828	34.5
6A	28	773	27.6	745	26.6	861	30.8
6B	46	886	19.3	809	17.6	874	19.0
6C	54	1291	23.9	1168	21.6	1330	24.6
5A	48	792	16.5	708	14.8	777	16.2
5B	23	299	13.0	309	13.4	294	12.8
5C	43	541	12.6	432	10.0	477	11.1
5D	46	687	14.9	638	13.9	724	15.7
4A	19	132	6.9	175	9.2	133	7.0
4B	22	274	12.4	380	17.3	282	12.8
4C	16	108	6.7	119	7.4	108	6.8
4D	27	332	12.3	350	13.0	316	11.7
4E	19	244	12.8	262	13.8	243	12.8
4F	21	285	13.6	286	13.6	295	14.1
4G	19	244	12.8	169	8.9	188	9.9
4H	18	173	9.6	188	10.4	169	9.4
3A	16	72	4.5	121	7.6	77	4.8
3B	17	191	11.2	236	13.9	196	11.5
3C	11	139	12.6	145	13.2	144	13.1
3D	28	120	4.3	139	5.0	114	4.1

[8] Available at http://www.capes.gov.br/ under Evaluation of Graduate Programs.

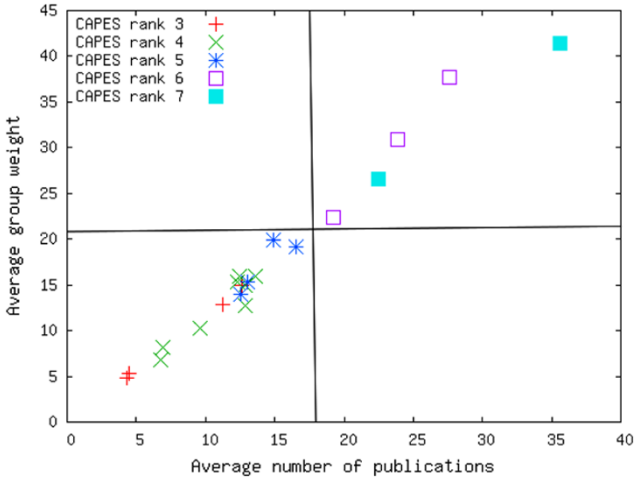


Fig. 11 Direct comparison between average number of publications and average group weight

6 Conclusion

This paper presents a study of structural properties of the global and Brazilian scientific collaboration networks in the area of Computer Science. In particular, we use *DBLP* (large database of mostly Computer Science publications) to characterize and compare these two networks, focusing on the intensity of relationships among individuals. We also develop a method to rank vertices in collaboration networks. The main idea of the proposed method is to rank individuals within a group (and not in absolute terms) using a metric that is proportional to the relationships of the individual with individuals outside the group. Intuitively, influential individuals within a group tend to play the role of information bridges between the group and the outside, and such metric attempts to identify them.

In order to evaluate the effectiveness of the proposed method, we compare the ranking of Computer Science researchers in Brazil according to different metrics with the ranking established by the Research Fellowship Program of CNPq. Our empirical results indicate that the proposed metric is the most effective when identifying the most influential researchers. In particular, the proposed metric exhibited higher precision and recall (and F-measure) than other classical metrics (including the state-of-the-art) when identifying 1A researchers.

We also show that our proposed method can be applied to rank groups of individuals. We apply our metric to groups formed by faculty members of Computer Science graduate programs in Brazil that have publications that appear in *DBLP*. We compare different metrics to rank the groups with the ranking of graduate programs established by CAPES. Our results indicate that our metric has a good correlation with the ranking of CAPES. In particular, it effectively identifies the most highly ranked Computer Science graduate programs in Brazil, indicating its suitability in correctly identifying influential groups in scientific collaboration networks.

Acknowledgements This work received financial support from CAPES (scholarship), and FAPERJ through the program "Jovem Cientista do Nosso Estado".

References

1. Barabási AL, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. Physica A 311(3–4):590–614
2. Figueiredo DR (2011) Introdução a redes complexas. In: de Souza AF Jr, Meira W (eds) Atualizações em Informática 2011, PUC-Rio, pp 303–358. Chap 7
3. Freire VP (2010) Uma métrica para ranqueamento em redes de colaboração baseada em intensidade de relacionamento. Master's thesis, Universidade Federal do Rio de Janeiro (UFRJ)/COPPE, in Portuguese
4. Gastwirth JL (1972) The estimation of the Lorenz curve and Gini index. Rev Econ Stat 54(3):306–316
5. Grossman J, Ion P (1995) On a portion of the well-known collaboration graph. Congr Numer 108:129–131, The Erdős Number Project, http://www.oakland.edu/enp/
6. Huang J, Zhuang Z, Li J, Giles CL (2008) Collaboration over time: characterizing and modeling network evolution. In: International conference on web search and web data mining (WSDM), pp 107–116
7. Laender AHF, de Lucena CJP, Maldonado JC, de Souza e Silva E, Ziviani N (2008) Assessing the research and education quality of the top Brazilian computer science graduate programs. SIGCSE Bulletin 40(2):135–145
8. Ley M (2009) Dblp: some lessons learned. Proc VLDB Endow 2(2):1493–1500
9. Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. J Am Soc Inf Sci Technol 58(7):1019–1031
10. Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge
11. Menezes GV, Ziviani N, Laender AHF, Almeida VAF (2009) A geographical analysis of knowledge production in computer science. In: 18th international conference on World wide web, pp 1041–1050
12. Nascimento MA, Sander J, Pound J (2003) Analysis of sigmod's co-authorship graph. SIGMOD Rec 32(3):8–10
13. Newman MEJ (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci USA 98(2):404–409
14. Newman MEJ (2004a) Coauthorship networks and patterns of scientific collaboration. Proc Natl Acad Sci USA 101(Suppl 1):5200–5205
15. Newman MEJ (2004b) Who is the best connected scientist? A study of scientific coauthorship networks. In: Complex Networks. Springer, Berlin, pp 337–370
16. Newman MEJ (2010) Networks: an introduction. Oxford University Press, Oxford
17. Onnela JP, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J, Barabási AL (2007) Structure and tie strengths in mobile communication networks. Proc Natl Acad Sci USA 104(18):7332–7336
18. Wagner C, Leydesdorff L (2005) Network structure, self-organization, and the growth of international collaboration in science. Res Policy 34(10):1608–1618