International Journal of Computer Science and Applications,
©Technomathematics Research Foundation
Vol. 9, No. 2, pp. 15 – 31, 2012

SCIENTIFIC COLLABORATION IN RESEARCH NETWORKS: A QUANTIFICATION METHOD BY USING GINI COEFFICIENT

GISELI RABELLO LOPES

Instituto de Informática, Universidade Federal do Rio Grande do Sul - UFRGS,
Porto Alegre, Rio Grande do Sul, Brazil,
grlopes@inf.ufrgs.br
http://www.inf.ufrgs.br/~grlopes

ROBERTO DA SILVA

Instituto de Física, Universidade Federal do Rio Grande do Sul - UFRGS,
Porto Alegre, Rio Grande do Sul, Brazil,
rdasilva@if.ufrgs.br
http://www.if.ufrgs.br/~rdasilva

MIRELLA M. MORO

Departamento de Ciência da Computação, Universidade Federal de Minas Gerais - UFMG,
Belo Horizonte, Minas Gerais, Brazil,
mirella@dcc.ufmg.br
http://www.dcc.ufmg.br/~mirella

JOSÉ PALAZZO MOREIRA DE OLIVEIRA

Instituto de Informática, Universidade Federal do Rio Grande do Sul - UFRGS,
Porto Alegre, Rio Grande do Sul, Brazil,
palazzo@inf.ufrgs.br
http://www.inf.ufrgs.br/~palazzo

In the scientific community, it is very common to try to create sound metrics for practically everything that can be measured. One of the current trends is to consider aspects from *social networks* for defining evaluation metrics. Following such a trend, our work proposes applying the Gini coefficient for evaluating research networks from two different perspectives. The first one analyzes the temporal evolution of research networks by considering the Gini coefficient of the distribution of researchers who have co-authored publications. The second one compares different internal collaboration networks of graduate programs and applies the Gini coefficient to support the ranking creation task. Both ideas are demonstrated through experiments that show the validity and applicability of our approach for quantifying scientific collaborations. Moreover, we also propose a new index that combines two metrics for evaluating the collaboration network of graduate programs. We believe that this index should be easily applied to other groups and research areas. This is the first time that the Gini coefficient is applied over social networks for evaluating research.

Keywords: Gini Coefficient; Social Networks Analysis; Co-Author Social Networks.

1. Introduction

The attempt to create metrics to analyze various issues from different perspectives has been a trend in the scientific community. Indeed, there are specialized events and publications on different perspectives of metrics, such as the *International Conference on Scientometrics and Informetrics* and the journal *Scientometrics*. With so much being researched and published, one current trend is to consider aspects from *social networks* for defining evaluation metrics.

Nonetheless, the interest of academia and the general public on social networks has grown rapidly. Specifically, the increasing interest in researching on Social Networks was encouraged by the popularization of online social networks, which are very rich and complex Web applications. Examples of such networks include *LinkedIn, Facebook* and *Google+,* among others, and each of those connects millions of users.

A related topic is Social Network Analysis (SNA), which assumes that the interacting units are the central point for the evaluation and analysis of social collaboration. The Social Network perspective includes theories, models and applications that are expressed in terms of relational concepts. Moreover, the rigorous measurement of the relationship defined by the linkages among the particles (nodes) is a fundamental component to infer properties of the studied network.

Some fundamental concepts used on SNA include actors and relational ties [Wasserman and Faust (1994)] [Knoke and Yang (2007)]. *Actors* are social entities that exhibit social linkages modeled by the social network. Actors are linked to other actors by *relational ties*. The range and the kind of these ties can be quite extensible. Thus, a critical and significant feature of a Social Network is the presence of relational information. The aforementioned growing interest and use of social network analysis has provided a consensus about the predominant principles of the networks that distinguish social networks analysis from analysis performed for a general kind of network. In addition to the use of relational concepts, Wasserman and Faust (1994) emphasize the following about social networks:

- Actors and their actions are regarded as correlated rather than uncorrelated;
- Relational ties between actors are viewed as channels for transferring resources;
- Network models focus on the individuals in the network structural environment, for providing opportunities for individual action;
- Network models conceptualize structure as lasting patterns of relations among actors.

Developing methods in the context of SNA is significantly important, because the analysis of a particular unit on a social network does not involve a single individual but an entity consisting of collections of individuals and links among them. Moreover, SNA is inherently an interdisciplinary effort: its concepts are developed in social theory, empirical research and formal mathematics and statistics [Wasserman and Faust (1994)]. Also, the pioneers in SNA come from sociology, social psychology and anthropology, and the first use of the term “social network” is granted to Barnes (1954).

The methods of SNA provide formal declarations about social processes and properties. Moreover, these concepts must be defined precisely and consistently. Once defined, these concepts can supply logical reasons about the social world [Freeman (1984)]. Recent works employed SNA to understand the interconnections, evolution and behaviors of whole researchers' communities [Ding (2011)][Menezes *et al.* (2009)] [Wang *et al.* (2010)].

Such a research (or scientific) social network represents scientific collaborations and relationships among peers of academic and industry networks. Here, actors represent authors and relational ties represent the relationships between pairs of authors. The relationship may be given by any kind of interaction between two researchers. For example, the presence of at least one co-authored paper between two researchers may determine a relational tie between them. For identifying such relations, we can use different data sources available at the web, such as DBLP, Google Scholar, CiteSeer, BDBComp, ISI-JCR, among others.

It is important to notice that joining the powerful analysis given by SNA and the available data about research communities, one can define metrics for evaluating the way research groups collaborate. Indeed, the pursuit of excellence in research areas and the competition for grants have motivated studies in research quality assessment using bibliometric measures [Egghe (2006)] [Egghe (2010)] [Hirsch (2005)] [Habibzadeh and Yadollahie (2008)] [Nicolaisen and Frandsen (2008)] [Ren and Taylor (2007)]. These metrics may be applied not only to assess research quality but also: to help the decision making of funding agencies, to assess quality of research-related features for budgetary allocation by governments, and to allocate human resources.

Furthermore, we can employ quality measures to define rankings, such as ranking journal and conference based on the quality of their editorial boards, ranking universities based on the quality of their researchers and faculty members, and ranking research project proposals based on the quality of their researchers' proponents. This is a very sensitive point as the development or the inhibition of a research group may be consequence of such an evaluation. Once again, there is an emphasis on employing bibliometric techniques in this context, especially citation statistics [Molinari and Molinari (2008)] [Ren and Taylor (2007)] [Silva *et al.* (2010)] [Yan and Lee (2007)]. However, a naïve application of bibliometrics measures may easily lead to an unfair rank.

Then, the contributions of this paper are as follows.

- We overview some central concepts of SNA and discuss the potential for using the Gini coefficient to quality assessment (Section 2);
- We introduce a new way of applying the Gini coefficient to SNA (Section 3);
- We experimentally show how to use the Gini coefficient to evaluate the evolution of a research social network (Section 4);
- We introduce the β-index, which is based on the Gini coefficient and experimentally employed for ranking Computer Science graduate programs (Section 5).

2. Related Work

This section overviews concepts related to Social Networks Analysis and describes the Gini coefficient and its potential to assess research quality.

2.1. *Social Networks Analysis Concepts*

SNA comprises two points of view: the *whole networks* and the *ego-centered* networks. The first one focuses on the *structural relationship* of the network with the social group. According to this view, networks are signatures of social identity, and the relationships patterns between individuals are represented by the preferences and characteristics mapping of the involved parts within the network [Watts (2003)]. The second (ego-centered) focuses on the social role of an individual, which can be understood not only by the groups (networks, circles) to which he belongs, but also by the position he holds in such networks. The differences between the two visions are covered in [Recuero (2005)].

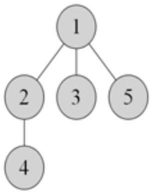
Nonetheless, a common way to represent a social network is through a graph *G=(N,E)*, in which nodes (or vertices) *n Є N*, and edges (or links, linkages, connections) *e Є E*. In the context of social networks, *N* is the set of network actors and *E* the set of relational ties among those actors. Some of the most common concepts used in social networks analysis are presented next [Freeman (1979)] [Hoser *et al.* (2006)] [Marsden (2002)].

Degree Centrality. The concept of degree centrality presumes that a node with many connections is considered important, whereas a node without connections is considered irrelevant. The degree centrality is calculated as the number of direct ties that involve a given node. If the network is an undirected graph (i.e. the connection between two nodes is undirected) the metric is just called *degree*. If the network is a directed graph (i.e. the connection between two nodes is directed) the metric is categorized into in-degree and out-degree, according to the direction of the relationships being considered.

Closeness Centrality. The concept of closeness centrality describes the level of freedom for the nodes in a network, or simply their capacity for independent action within the network. The closeness centrality of a node is calculated as the inverse of the sum of the shortest paths (geodesic distance) between a particular node and all other nodes.

Betweenness Centrality. The concept of betweenness centrality describes the global location of a node in the network. This concept describes the intermediary location of a node among indirect relationships linking other nodes. The betweenness centrality of a node is calculated by the number of shortest paths between any two nodes that pass through the specified node (it can be normalized or not).

Density. The density of a network is defined based on the degree centrality. The density is calculated as the number of edges divided by the number of all possible edges of this network. The number of all possible edges changes according to the kind of graph describing the network. If the network is an undirected graph, the possible number of



(a)

Measure	Value
Degree Centrality (#1)	3
Closeness Centrality (#1)	0.2
Betweenness Centrality (#1)	5
Density	0.4
Diameter	3

(b)

Fig. 1. Example of a simple social network (a) and its metric values (b).

connections between each two nodes is 1. If the network is a directed graph, the possible number of connections between each two nodes is 2. A network entirely connected has the density of 1. According to Hoser *et al.* (2006), this concept is not useful when multiple edges are allowed or when the edges are weighted, because it is not possible to define the total number of possible connections.

Diameter. The diameter is a metric associated to the graph distance. This metric is calculated as the maximum value obtained among all shortest path evaluated between two nodes of the network graph (i.e., the longest distance between any pair of nodes belonging to the graph). In social networks, this value usually is small. The diameter and the density are metrics commonly used for comparing social networks.

Figure 1 shows an example for all metrics. Specifically, Figure 1a illustrates a simple undirect SN composed of five actors and Figure 1b shows the measures estimated for this SN. In this example, the network has a unique connected component. When there is no such a component, the main component (connected component with the largest number of nodes) must be used to the measures estimation. The first three metrics are calculated for specific nodes. Here, the chosen node is number 1 (#1). The degree centrality for node #1 is 3, referring to the number of its relational ties (edges). The value of closeness centrality is 0.2, and it was calculated as 1 divided by the sum of shortest path among node #1 and all others, i.e. (1/(3 shortest path of length 1 + 1 shortest path of length 2)). The betweenness centrality (non-normalized) is 5, because there are 5 shortest paths among pairs of nodes passing through node #1 (i.e., #2 and #3, #2 and #5, #3 and #4, #3 and #5, and #4 and #5). The two last measures are estimated for the SN as a whole. The density value is 0.4, i.e., the total number of edges existent (4) divided by the total number of possible connections (10). Finally, the diameter value is 3, i.e., the value of maximum shortest path (between the nodes #3 and #4, or #4 and #5).

In the next section, we describe the Gini coefficient, which we propose as a new metric to describe and quantify social networks.

2.2. *Gini Coefficient*

The Gini coefficient is a measure of statistical dispersion proposed in 1912 by the Italian statistician Corrado Gini in his paper “Variability and Mutability” [Gini (1955)]. This coefficient is commonly used to evaluate the inequality of income and wealth distributions but can be used for other distributions [Subramanian and Kawachi (2004)].

The Gini coefficient is defined based on the Lorenz curve. The Lorenz curve was created by Max O. Lorenz in 1905. The Lorenz curve is a graph that represents the cumulative distribution of a probability density function. Such a function is built as a ranking of the members of the population disposed in increasing order of wealth (or whatever amount is studied). Let us denote these amounts as $h_1 < h_2 < ..., h_{n-1} < h_n$ and consider Equation 1, which calculates the fraction of wealth (cumulative distribution) corresponding to the fraction of people $f_i = i/n, i = 1, ..., n$ in the population.

$$\Phi(h_i) = \frac{1}{\left(\sum_{j=1}^{n} h_j\right)} \sum_{k=1}^{i} h_k \tag{1}$$

In recent experiments, we have employed Lorenz curves to study the distribution of h-index of researchers in conferences [Silva *et al.* (2010)] [Silva *et. al.* (2011)]. Figure 2 illustrates the typical plots for different conferences analyzed in that work.

In Figure 2, the percentage of individuals is plotted on the x-axis and the percentage of the variable values (in the original conception this variable is the wealth of population) on the y-axis. The distribution is perfectly equalitarian when every individual has the same variable value. In this case, the bottom “N”% of the group of individuals always has “N”% of the variable value, and its curve is “y=x”, which is called the *perfect equality line*. On the other hand, the perfectly unequal distribution is the one in which only one individual has all the variable value. Finally, the curve is “y=0” for all “x<100%”, and “y=100%” when “x=100%”, which is called the *perfect inequality line*.

The Gini coefficient is calculated as the area between the *perfect equality line* and the observed Lorenz curve, as defined by Equation 2.

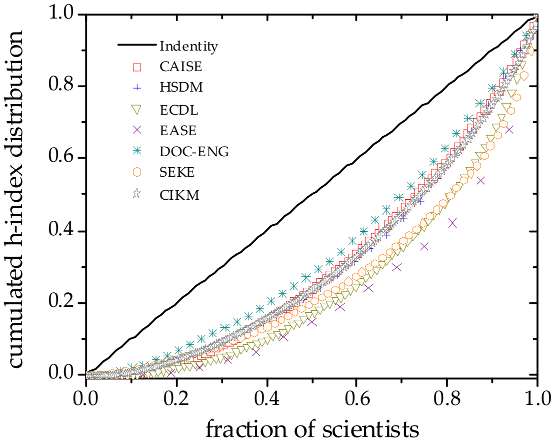


Fig. 2. Lorenz curves for the h-index distributions of researches in conferences of Engineering Software.

$$g = 1 - 2\int_0^1 \Phi(h)dh \tag{2}$$

which is numerically approximated by a trapezoidal formula, leading to Equation 3.

$$g \cong 1 - \frac{\Phi(h_0) + \Phi(h_n)}{n} - \frac{2}{n}\sum_{k=1}^{n-1}\Phi(h_k) = 1 - \frac{1}{n}\sum_{k=1}^{n}[\Phi(h_k) + \Phi(h_{k-1})] \tag{3}$$

where $\Phi(h_0) = 0$ and $\Phi(h_n) = 1$ for construction. The Gini coefficient represents the percentage of the area between the *perfect equality line* and the *perfect inequality line*. This coefficient is directly proportional to the inequality of the distribution. The interval of the Gini coefficient result varies from 0 to 1 (or 0% to 100%). A low value of Gini coefficient indicates a more equal distribution among the parts, and a high value indicates a more unequal distribution. Our hypothesis is that such an amount is also appropriated to quantify whether the research collaboration is equalitarian in social networks, as described in the next section.

Regarding the current uses of the Gini coefficient, Pissard and Prieur (2007) have proposed techniques to quantify parameters in social networks using it. They consider the context of users in a social network uploading photos. Our work considers a complete different scenario, in which we want to quantify scientific collaboration in research social networks.

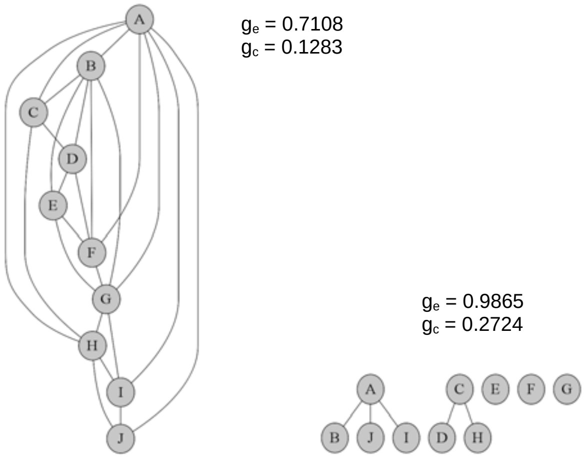
In another related work, Silva *et al.* (2010) applied Gini coefficient to analyze *h*-index distributions for ranking program committees of conferences and editorial boards of journals. However, the authors do not apply the coefficient to analyze the Social Networks connections, but to analyze the distribution of *h*-index, exploring the facet of bibliometric citation.

Here, differently from previous work, we apply Gini coefficient to analyze research social networks (Section 3). Furthermore, we employ it in two different perspectives: to analyze how a network evolves in terms of productivity (Section 4) and to evaluate research groups (Section 5).

3. Gini Coefficient applied on Social Networks Analysis

In this paper, we propose that the Gini coefficient be used as a metric (as those presented in Section 2) within Social Networks Analysis. We do so in two forms: g_e and g_c. The difference between them is the distribution adopted for the social network analysis from which the Gini coefficient is calculated (as presented in the Section 2.2).

The first one, named g_e, defines the Gini coefficient considering that the relational ties (edges) between actors (nodes) can be viewed as a distribution of the possible pairs of actors within the network (each actor combined with all of its neighbors in the SN). For each pair of actors, the value associated is either 0 (zero) when there is no relationship between them, or the normalized weight of their relationship (1 in unweighted SN). The g_e distribution is perfectly equalitarian when every pair of actors



(a) (b)

Fig. 3. Examples of Social Networks: (a) connected network and (b) poorly connected network.

has the same weight. On the other hand, it is completely nonequalitarian when only one pair of actors has a relationship. A low value of Gini coefficient indicates a more equal distribution (i.e. a SN more connected), and a high value indicates a more unequal distribution (i.e a SN more disconnected).

The second one, named g_c, defines the Gini coefficient based on the grade of connectivity of the network actors. In this case, we are interested in measuring the distribution of relationships among actors. Then, the distribution is formed by the number of connections of each actor in the SN. Complementary to g_e, high values of Gini coefficient indicates that the connectivity of the SN is homogeneous, whereas low values reveal that the SN connectivity is not homogeneous (i.e., the connections only occur among few actors, whereas the majority is not connected).

For example, Figure 3a and 3b present two distinct social networks for a group of actors. Their g_e are 0.7108 and 0.9865, and g_c are 0.1283 and 0.2724. As one can see, the left-side network is very well connected and its relations are very well distributed. On the other hand, the right-side network is poorly connected, with weak relations within the whole network. It is important to notice that we are emphasizing how the Gini coefficient values do indeed characterize the relationship distributions of those networks.

4. Temporal Evolution of a Social Network with Gini Coefficient

As stated in Section 4, we propose two different ways to apply Gini coefficient in Social Networks Analysis. In this section, we apply them to a Co-Authorship Social Network, in which actors are researchers and relational ties are research collaboration (expressed through co-authored papers) between pairs of researchers. Furthermore, we considered the social network as a weighted graph in which the weights are given by Equation 4.

$$w_{ij} = \frac{n_{ij}}{n_i} \tag{4}$$

where n_{ij} denotes the number of common papers between the pair of the neighbors $<i,j>$. Here $n_i = \sum_{k=1}^{m} n_{ik}$ (i.e., the number of papers authored by researcher i).

It is also important to notice that the weights of each edge are non-symmetric, because w_{ij} is different of w_{ji} for $n_i \neq n_j$. For example, given a senior researcher *r* and one of her students *s*, the weight from *r* to *s* is most probably smaller than from *s* to *r*, because *r* has authored more papers with other people than *s*. Next, we present the dataset used for building the social networks in this experiment and the respective analysis of Gini coefficient.

4.1. *Dataset Description*

In this experiment, we used a network composed by 27 researchers involved in the INWeb, the Brazilian National Institute of Science and Technology for the Web (http://inweb.org.br). This project began in 2008 and all of its researchers are faculty members (including full, associate and assistant professors) in the participant Brazilian institutions, namely: UFMG, UFRGS, UFAM and CEFET-MG (all with graduate programs in Computer Science).

The dataset used to build the networks was obtained from DBLP on August 03, 2010. In order to evaluate the evolution of the social networks, we have considered two networks built using different time intervals. The first time interval includes the publications of its researchers until 2007, defining a network called *SN2007*. The second time interval includes publications until 2010, defining a network called *SN2010*. The project started in 2008, so *SN2007* reflects the researchers' collaborations before INWeb begun, whereas *SN2010* shows the researchers' collaborations until two years after its beginning (before and during the development of the project). Now we discuss how the Gini coefficient can be used for evaluating the impact of having the project in the collaborations among its researchers.

Figure 4 shows the evolution occurred in the researchers' collaborations considering a comparative between SN2007 and SN2010. Instead of presenting one figure for each network, we have assembled both in Figure 4, as follows. The numbers within the nodes specify the researchers: the first number (#Id) identifies the institution affiliation of the researchers according to Table 1, and the second number identifies each researcher within that institution. The researchers that co-authored at least one paper are connected by edges: the solid gray lines represent connections that are not intensified during the INWeb project; the solid black lines represent the connections intensified during the development of the project (from 2008 to 2010); and the dashed lines represent the new connections that occurred during the project until year 2010.

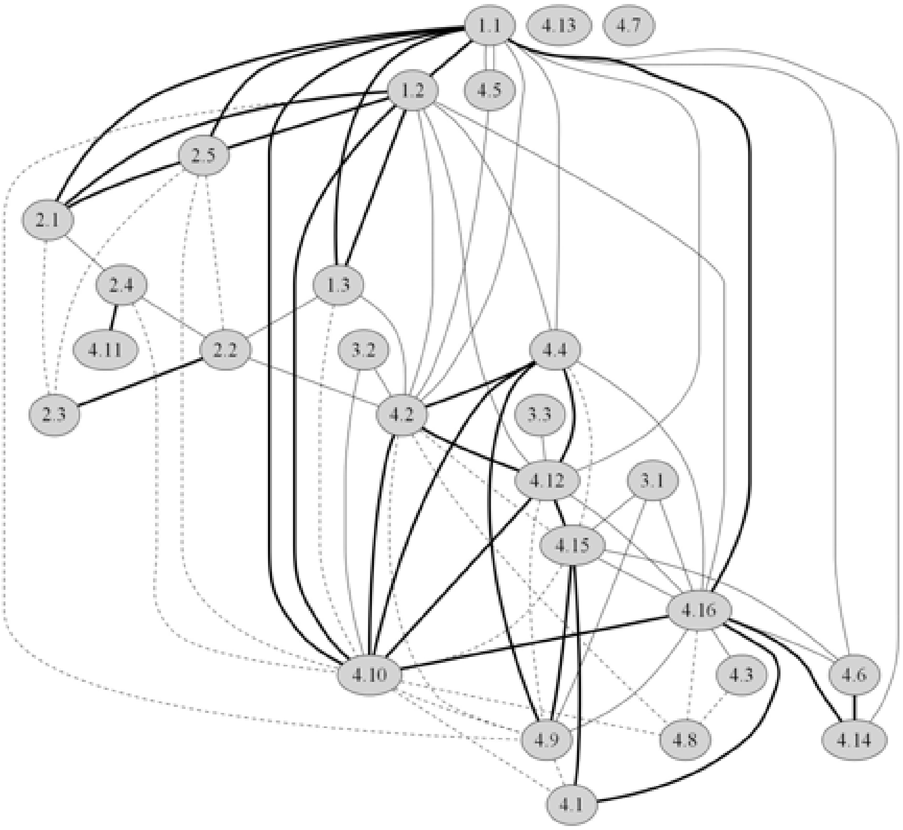


Fig. 4. Comparative between Social Network of INWeb before the project's beginning and during is development: gray lines for non-intensified connections, black lines for intensified connections, and dashed lines for new connections.

Table 1. The institutions participating in the INWeb project

#Id	Institutions
1	UFAM, Manaus, AM, Brazil
2	UFRGS, Porto Alegre, RS, Brazil
3	CEFET-MG, Belo Horizonte, MG, Brazil
4	UFMG, Belo Horizonte, MG, Brazil

4.2. *Evaluations*

This section evaluates SN2007 and SN2010 from different perspectives: global analysis, cooperative analysis and Lorenz curves.

4.2.1. *Global Analysis*

The first analysis considers the distribution of weights of all possible relational ties between researchers, including the weight zero between researchers that have not co-authored any paper. This analysis of the Gini coefficient among all pairs of researchers is important to evaluate the global collaboration between the researchers of the studied SN.

The results of this first analysis (see Analysis 1 at Table 2) show that the value of Gini coefficient is lower in SN2010 (*SN2010_allPairs*) than in SN2007 (*SN2007_allPairs*). These results are coherent and they are justified as SN2010 is more connected than SN2007. However, the difference between them is very small and statistically insignificant. It is possible to notice that high values of Gini coefficient were obtained for both networks. These results indicate unequal distributions, *i.e.* this collaboration network of INWeb is still very disconnected (in relation to one totally connected network) in both considered time intervals.

Table 2. Gini Coefficient Values.

Analysis	Distribution	Gini Coefficient
(1)	SN2007_allPairs SN2010_allPairs	0.9471 0.9327
(2.1)	SN2007_Pairs2007 SN2010_Pairs2007	0.5824 0.5735
(2.2)	SN2007_Pairs2010 SN2010_Pairs2010	0.7009 0.6160

The INWeb research project has duration of five years and is still in development. At the end of the five years, the objective is to have an increasing in the cooperation pattern among inter-institutional researchers. From the social point of view, this type of analysis is fundamental to understand the characteristics and properties of the SN to be studied.

4.2.2. *Cooperative Analysis*

The second analysis considers only the pairs of researchers that have co-authored paper (weights nonzero) in one of the analyzed SNs (SN2007 or SN2010). This analysis of the Gini coefficient considering only the cooperative pairs of researchers is important to evaluate the genuine level of collaboration between those pairs that have relational ties modeled by the SN.

Cooperative Pairs until 2007. The results of analysis considering only the cooperative pairs of researchers until 2007 (see Analysis 2.1 at Table 2) show that the value of Gini coefficient is lower in SN2010 (*SN2010_Pairs2007*) than in SN2007 (*SN2007_Pairs2007*). This difference reflects the improvement in the homogeneity of weight distributions among the researchers that have already collaborated before the beginning of the INWeb project. The collaborations between the researchers that are intensified, as can be observed in Figure 4, contributed to a more equalitarian distribution in SN2010 than in SN2007.

Cooperative Pairs until 2010. The results of analysis considering only the cooperative pairs of researchers until year 2010 (see Analysis 2.2 at Table 2) show that the value of Gini coefficient is lower in SN2010 (*SN2010_Pairs2010*) than in SN2007 (*SN2007_Pairs2010*). Once more, this difference reflects the improvement occurred in the collaborations of the researchers after the beginning of the INWeb project. This occurs as the pairs of authors considered in both SNs are equal. It is possible to notice that the collaborations between the researchers were intensified, as may be observed in Figure 4, and the value of Gini coefficient identified this intensification of the connectivity. Moreover, it is also possible to notice that high values were obtained for both networks. These results indicate unequal, but better distributions in both SNs. In summary, it is possible to affirm that: (i) in SN2007, there are pairs of researchers that are not connected and the distribution of weights is nonequalitarian, and (ii) in SN2010, some already existent collaborations are intensified (increased weights) whereas new collaborations emerged (first publications co-authored, probably low weights), making the distribution of weights nonequalitarian but better in values.

4.2.3. *Lorenz Curves*

As stated before, the Gini coefficient is calculated as the area between the perfect equality line (45 degree slope) and the observed Lorenz curve. The graphical representations of the Lorenz Curves of the distributions analyzed in this case study are presented in Figure 5. In this figure, the cumulative percentage of pairs of researchers is

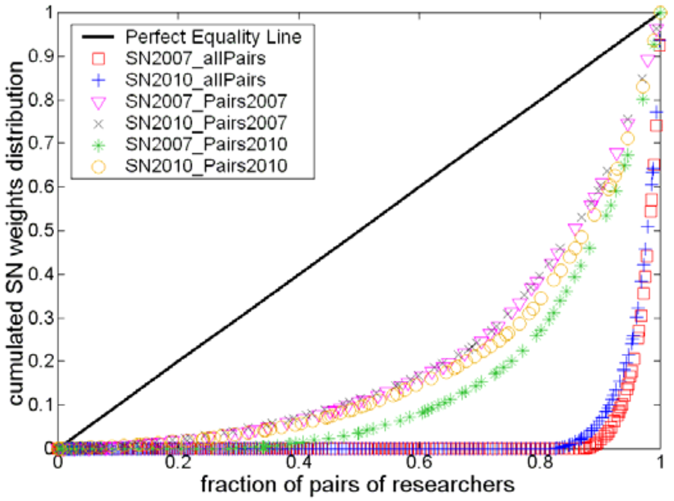


Fig. 5. Lorenz Curves for the SNs distributions.

plotted on the x-axis; the cumulative percentage of the relationship weights, on the y-axis. The legend of the distributions is the same used in Table 2.

In the second analysis (cooperative), only the pairs of researchers that have some relational tie between them were selected. In this case, it is possible to notice that there exist lower areas between the Lorenz curves of the distributions and the perfect equality line, different from the results with the areas corresponding to the first analysis (global). In other words, these values indicate more equal distributions than in the first analysis, i.e. SNs more connected. Such results are coherent by the data that were analyzed.

As a summary of this first set of experiments, we have employed the Gini coefficient to analyze a real research social network. The results show significant evidence of the validity and applicability of the Gini coefficient on the context of Social Networks. This approach is a promising method for a post evaluation of cooperative and of mobility research projects from the point of view of scientific cooperation analysis.

5. Ranking research groups with Gini coefficient

This next experiment applies the Gini coefficient to comparatives analysis among different Research Social Networks, in which each network has probably different size because it is composed by different sets of researchers. Specifically, we propose to rank research groups by evaluating the collaborative behavior among the researchers that belong to their social networks. Here we show that the Gini coefficient can be employed to perform such assessment combined with another metric as defined next.

Average of co-authored papers. This measure calculates the average of number of papers co-authored by pairs of researchers, as defined by Equation 5.

$$\rho = \frac{\sum_{i=1}^{m} \sum_{j=i+1}^{m} n_{ij}}{e} \tag{5}$$

where n_{ij} denotes the number of common papers between the pair of research neighbors $<i,j>$, m denotes the total number of authors, and e denotes the total number of edges existent in the SN (i.e., the SN considers only pairs of researchers $<i,j>$ that have at least one co-authored paper).

β-index. In order to fairly evaluate the networks, we propose a new metric for the average number of papers co-authored by pairs of researchers that cooperated together, called *β-index*. This metric considers the Gini coefficient evaluation g_c (i.e., the distribution of number of co-authors of each researcher in a SN) and the average of co-authored papers, as defined by Equation 6.

$$\beta = \frac{\rho}{g_c} \tag{6}$$

where ρ is the average of co-authored papers of the SN and g_c is the Gini coefficient considering the distribution of connectivity of this same SN.

For ranking purposes, our hypothesis is that high quality research groups have a high average of co-authored papers and, simultaneously, their authors have similar behavior in relation to connectivity (low Gini coefficient). This behavior is reflected by high values of *β-index*.

Next, we present the dataset used in this case study and evaluate the results of the analysis of the social networks considering graduate research programs.

5.1. *Dataset Description*

In order to define the dataset for our experimental evaluation, the first challenge was to choose a set of research groups to be ranked. The second challenge was to define a baseline rank to which ours could be compared against. We have solved both challenges by selecting a set of Brazilian graduate programs, for which a federal agency defines an official rank at each three years.

Specifically, the dataset includes 732 researchers from 27 Brazilian graduate programs in Computer Science and their publications (extracted from DBLP in August 2010). In Brazil, all graduate programs are evaluated each year by CAPES (the federal agency for qualification of human resources). Then, for each three years of evaluation, the graduate programs are classified in a *Likert* scale of seven levels (from 1 to 7). This classification is performed by experts and is based on a series of quality criteria. The top quality levels are 7 and 6, and they represent the graduate programs with performance comparable to top international programs. In this experiment, the set of programs includes all those from levels 7, 6, and 5, and a selected set of levels 4 and 3 (below level 3, the programs need to promote a restructuration or will be closed). For public transparency, all the data are disclosed and available at the CAPES website. This data set is presented in Table 3, in which the first column corresponds to the selected graduate program name and the second column presents the respective CAPES evaluation.

5.2. *Evaluations*

In order to evaluate the graduate programs, we considered the CAPES evaluation as baseline, and we compare it against our new measure (*β-index*) by using the Spearman's coefficient, a common measure that enables to determine the ranking correlation. Since CAPES ranks the programs by levels, all programs in the same level can be considered as tied. Therefore, we use the variation of Spearman’s coefficient that deals with tied ranks [Siegel and Castellan (1988)]. For this coefficient, the higher it is, the higher the correlation between the rankings being compared. The significance level of the obtained results is also evaluated. The statistical significance threshold of 0.01 is used and the results are summarized in Table 4. This table presents the results in ascending order according to the Spearman's coefficient, i.e. ordered from the worst to the best result.

Table 3. Selected set of graduate programs in Computer Science and their respective CAPES evaluations (according to the tri-annual ranking 2007-2009, published in 2010).

Graduate program	CAPES evaluation
COPPE/UFRJ	7
PUC/RIO	7
UFMG	7
UFPE	6
UFRGS	6
UNICAMP	6
USP/SC	6
UFF	5
USP	5
PUC/PR	4
PUC/RS	4
UFAM	4
UFBA	4
UFC	4
UFCG	4
UFES	4
UFPR	4
UFRJ	4
UFRN	4
UFSC	4
UFSCAR	4
UNB	4
UNISINOS	4
PUC/MG	3
UCPEL	3
UFG	3

Table 4. Spearman’s Coefficient values comparing rankings generated using different measures with baseline (CAPES evaluation).

Measure	Ordering	Spearman’s Coefficient	Significance Level
Average (ρ)	Descending	0.345283448380	Not correlated
Gini (g_c)	Ascending	0.421609263285	Not correlated
β-index	Descending	0.641665508855	Correlated

As the table shows, using the Average (ρ) and the Gini coefficient (g_c) measures alone does not provide good results. It is important to notice that ρ measures the intensity of the existent relationships. The results show that only this measure is not sufficient to order correctly the graduate programs, considering the CAPES baseline. Furthermore, using the Gini coefficient alone was also not sufficient to this purpose. The results of Spearman coefficient show that the rankings generated by both are not correlated with the baseline from CAPES.

However, as the results of Spearman coefficient stated, when the two measures (Average and Gini) were combined into *β-index*, the ranking generated for graduate programs was correlated to the CAPES evaluation. These results show evidences that the best graduate programs are those in which their researchers have a collaborative behavior with both: high intensity (high ρ) and more homogeneous distribution (low g_c), i.e., high *β-index* values. These results corroborate our hypothesis that the expected behavior of high quality research groups is having high Average of co-authored papers with the majority of researchers having the same behavior in relation to the connectivity (low Gini coefficient).

6. Concluding Remarks

The main novelty of this paper is to present a new use of Gini Coefficient for analysis of scientist groups as research Social Networks. Specifically, we have proposed two different applications of Gini Coefficient summarized as follows.

The first contribution is to use Gini Coefficient to evaluate weighted networks. The metrics commonly used for Social Networks Analysis do not consider the weights of relationships (if they exist) between the actors. On the other hand, we have employed a weighted graph where the weight of each edge is the number of common papers normalized for all contributions of the network. Then, we have successfully employed the Gini coefficient to measure the homogeneity level of collaboration.

The second contribution is to use of the Gini Coefficient combined with average of collaborations for ranking research groups based on the collaborative behavior of their researchers. The hypothesis was that the best groups (research programs in our experimental evaluation) are those with most of the researchers contributing to the group network with collaborations, whereas the worst ones are those with only few researchers with good level of collaboration. We have also shown that even the measures of Average (ρ) and Gini (g_c) alone are not sufficient to rank research groups. However, when combined, they can be indeed used as an assessment criterion for ranking purposes.

The initial results pointed out indices of the validity and utility of our approach. As future work, we plan to expand this study to other areas of knowledge and consider other datasets other than DBLP.

Acknowledgments

This research is partially supported by project grant 57.3871/2008-6 (INWeb), and by the authors' scholarships and individual research grants from CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brazil.

References

Barnes, J. (1954): Class and Committees in a Norwegian Island Parish. Human Relations, v. 7, n. 1, pp. 39–58.

Ding, Y. (2011): Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. Journal of Informetrics, v. 5, n. 1, pp. 187–203.

Egghe, L. (2006): An improvement of the h-index: the g-index. ISSI Newsletter, v. 2, n. 1, pp. 8–9.

Egghe, L. (2010): The hirsch-index and related impact measures. Annual Review of Information Science and Technology, v. 44, pp. 65–114.

Freeman, L. (1979): Centrality in social networks: I. Conceptual clarification. Social Networks, v. 1, n. 3, pp. 215–239.

Freeman, L. (1984): The impact of computer based communication on the social structure of an emerging scientific specialty. Social Networks, v. 6, pp. 201–221.

Gini, C. (1955): Variabilità e mutabilità, 1912, Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi.

Subramanian, S. V. and Kawachi, I. (2004): Income inequality and health: What have we learned so far? Epidemiol Rev, v. 26, n. 1, pp. 78–91.

Habibzadeh, F. and Yadollahie, M. (2008): Journal weighted impact factor: A proposal. Journal of Informetrics, v. 2, n. 2, pp. 164–172.

Hirsch, J. E. (2005): An index to quantify an individual's scientific research output. Proc. Of the National Academy of Sciences, v. 102, n. 46, pp. 16 569–16 572.

Hoser, B. *et al.* (2006): Semantic Network Analysis of Ontologies. European Semantic Web Conference - ESWC 2006, Lecture Notes in Computer Science, v. 4011, pp. 514–529.

Knoke, D.; Yang, S. (2007). *Social Network Analysis*, 2.ed. Sage Publications, Inc. (Quantitative Applications in the Social Sciences).

Marsden, P. (2002): Egocentric and sociocentric measures of network centrality. Social Networks, v. 24, n. 4, pp. 407-422.

Menezes, G. *et al.* (2009): A geographical analysis of knowledge production in computer science. A geographical analysis of knowledge production in computer science. Proc. of the 18th International Conference on World Wide Web (WWW '09). Madrid, Spain, pp. 1041–1050.

Molinari, J.-F. and Molinari, A. (2008): A new methodology for ranking scientific institutions. Scientometrics, v. 75, n. 1, pp. 163–174.

Nicolaisen, J. and Frandsen, T. F. (2008): The reference return ratio. Journal of Informetrics, v. 2, n. 2, pp. 128–135.

Pissard, N.; Prieur, P. (2007): Thematic vs. social networks in web 2.0 communities: A case study on Flickr groups. Proc. of Algotel Conference. Ile d'Oléron, France, pp. 31-34.

Recuero, R. (2005): Redes Sociais na Internet: considerações iniciais. E-Compós, Brasília, v.2.

Ren, J. and Taylor, R. N. (2007): Automatic and versatile publications ranking for research institutions and scholars. Commun. ACM, v. 50, n. 6, pp. 81–85.

Siegel, S.; Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. 2nd ed. New York: McGraw-Hill.

Silva, R. *et al.* (2010): Statistics for Ranking Program Committees and Editorial Boards. http://arxiv.org/abs/1002.1060.

Silva, R. *et al.* (2011): Universality in Bibliometrics. Physica A, v. 391, n. 5, pp. 2119–2128.

Wang, C. *et al.* (2010): Mining advisor-advisee relationships from research publication networks. Proc. Of ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, Washington, DC, USA, pp. 203–212.

Wasserman, S.; Faust, K. (1994). *Social Network Analysis: methods and applications.* Cambridge University Press.

Watts, D. J. (2003). *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company, February.

Yan, S. and Lee, D. (2007): Toward alternative measures for ranking venues: a case of database research community. Procs. of the 7th ACM/IEEE-CS joint conference on Digital libraries (JCDL '07). Vancouver, BC, Canada, pp. 235–244.