

Applying Gini Coefficient to quantify Scientific Collaboration in Researchers Network

Giseli Rabello Lopes

Instituto de Informática

Universidade Federal do Rio Grande
do Sul - UFRGS
Porto Alegre, Brazil

grlopes@inf.ufrgs.br

Roberto da Silva

Instituto de Informática

Universidade Federal do Rio Grande
do Sul - UFRGS
Porto Alegre, Brazil

rdasilva@inf.ufrgs.br

J. Palazzo M. de Oliveira

Instituto de Informática

Universidade Federal do Rio Grande
do Sul - UFRGS
Porto Alegre, Brazil

palazzo@inf.ufrgs.br

ABSTRACT

Some of the metrics more commonly used for Social Networks Analysis (SNA) do not consider the weights of relationships between the actors of the analyzed Social Network, if it exists. These weights aim to measure the importance of the relational ties between actors and also are important to be considered in a SNA. This paper purposes the Gini Coefficient to be applied on Social Networks Analysis. Our initial results demonstrate the validity and applicability of this approach for a collaboration of Brazilian network scientists.

Categories and Subject Descriptors

G.3 [Mathematics of Computing]: Probability and Statistics.

J.4 [Computer Applications]: Social and Behavioral Sciences.

General Terms

Measurement.

Keywords

Gini Coefficient, Social Networks Analysis, Co-Author Social Networks

1. INTRODUCTION

The Social Network Analysis (SNA) is based on the assumption that the importance of relationships among the interaction units is a central point to the evaluation and analysis of social interaction. The Social Network perspective includes theories, models, and applications that are expressed in terms of relational concepts. Thus, to measure 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 [1, 2]. *Actors* are social entities that have 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 defining feature of a Social Network is the presence of relational information.

The growing interest and use of social network analysis has provided a consensus about the central principles of the networks perspective that distinguish social networks analysis from other analysis performed specifically for other kind of networks.

In addition to the use of relational concepts, Wasserman and Faust [1] emphasize the following:

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

The interest of academics and of the general publics on social networks has grown rapidly. The increasing interest in researching on Social Networks was encouraged by the popularization of online social networks, which are very interesting Web applications. Examples of online social networks include *MySpace*, *Facebook*, *Bebo*, *Cyworld*, *Skyblog*, *Hi5*, *Friendster*, *Orkut*, *Live Journal*, among others. Nowadays, this kind of network is commonly used, and each network connects millions of users.

Another example of the social networks is a co-authorship social network which represents the relationships of scientific collaboration. In this case, actors represent authors and relational ties represent the relationships between pairs of authors. The presence of at least one co-authored paper between two authors determines a relational tie between them. Some examples of data sources for the construction of scientific collaboration networks are DBLP (*Digital Bibliography & Library Project*), Google Scholar, CiteSeer, BDBComp, ISI-JCR, among others.

The critical importance of developing methods in the context of social network analysis is the fact that the analysis of a particular unit on a social network is not a single individual but an entity consisting in collections of individuals and links among them. The social network analysis is inherently an interdisciplinary effort. The concepts of social network analysis are developed as part of advances in social theory, empirical research and formal mathematics and statistics [1]. The pioneers in social network analysis come from sociology, social psychology and anthropology, and the first use of the term “social network” is awarded to Barnes [3].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WIMS'11, May 25-27, 2011 Sogndal, Norway
Copyright © 2011 ACM 978-1-4503-0148-0/11/05... \$10.00.

The methods of social network analysis 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 [4]. Recent works employed social network analysis to understand the interconnections, evolution and behaviors of researchers' communities [5, 6, 7].

The social networks analysis came from two great points of view: the *whole networks* and *ego-centered networks*. In the first vision, the focus is 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 mapping the preferences and characteristics of the involved in the network [8]. In the second vision, the focus would be on the social role of an individual, which could be understood not only by the groups (networks) to which he belongs, but also by the position he holds in such networks. The difference between the two visions is the corpus of analysis chosen by the researcher [9].

The paper is organized as follows. In section II we describe some known properties that can be calculated in the graph defined according to the studied social network. In section III, we define the Gini coefficient as well as in section IV we show as it can be determined in social networks. The section V illustrates a study case while in section VI we finally present our main conclusions and summaries..

2. SOME BASIC APPROACHES TO SOCIAL NETWORKS ANALYSIS

The basic approach to analyze a network is to represent it as a graph $G:=(N,E)$, with nodes (vertices) $n \in N$ and edges (links, linkages, connections) $e \in E$. In the context of social networks, the nodes of this graph represent the network actors and the edges represent the relational ties among the actors of this network. Some of the basic approaches used in social networks analysis are presented below [10, 11, 12].

2.1 Degree Centrality

The concept of *degree centrality* presumes that a node that has many connections is considered important, while 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 no directional, the metric is just called *degree*. If the network is an directed graph, i.e. the connection between two nodes is directional, the metric is categorized into *in-degree* and *outdegree* according to the direction of the relationships being analyzed.

2.2 Closeness Centrality

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

2.3 Betweenness Centrality

The concept of *betweenness centrality* describes the global location of a node in the network. This approach reflects the intermediary location of a node along indirect relationships linking other nodes. The *betweenness centrality* of a node is

calculated by the normalized number of shortest paths between any two nodes that pass through the desired node.

2.4 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 will change according to the kind of graph describing the network. If the network is an undirected graph, the possible number of 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. [10], this concept is not useful when multiple edges are allowed or when the edges are weighted, because no total number of possible connections can be evaluated.

2.5 Diameter

The *diameter* is another metric associated to the connection degree of a graph. This metric is calculated as the maximum value obtained among all shortest path evaluated between two nodes of the network graph. In social networks, this value usually is small. The *diameter* as well as *density* are metrics usually used for comparing social networks.

In the next section, we define the Gini coefficient our new purpose to describe and quantify social networks.

3. 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" [13]. This coefficient is commonly used to evaluate the inequality of income and wealth distributions but can be used for other distributions.

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 graphical that represents the cumulative distribution of a probability density function, that is constructed by following way: a ranking of the members of the population must be disposed in increasing order of wealth (or whatever amount that we wish to study). Let us denote these amounts as $h_1 < h_2 < \dots, h_{n-1} < h_n$. Let us consider equation 1 which calculates the fraction of wealth (cumulative distribution) corresponding to the fraction of

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

people $f_i = i/n$, $i = 1, \dots, n$ in the population. In recent experiments, we have calculated the Lorenz curves to study the distribution of h-index of researchers in conferences (see ref. [14]). The typical plots of Lorenz curves for the different analyzed conferences in this reference can be observed in Figure 1.

In this graphical, the percentage of individuals being analyzed by distribution is plotted on the x-axis; the percentage of the variable values (in the original conception this variable is the wealth of population) of the distribution being analyzed, on the y-axis. The distribution will be perfectly equalitarian when every individual

has the same variable value. In this case, the bottom “N”% of the group of individuals would always have “N”% of the variable value. In the graph, the curve represented will be “y=x”; called the *perfect equality line*. On the other hand, the perfectly unequal distribution will be one in which only one individual has all the variable value. In the graph, the curve represented will be “y=0” for all “x<100%”, and “y=100%” when “x=100%”; called the *perfect inequality line*.

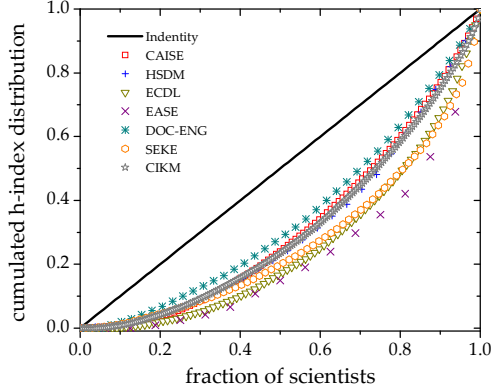


Figure 1. Lorenz curves for the h-index distributions of researches in conferences of Engineering Software.

The Gini coefficient is calculated as the area between the *perfect equality line* and the observed Lorenz curve, i.e.,

$$g = 1 - 2 \int_0^1 \Phi(h) dh \quad (2)$$

that numerically is approximated by a trapezoidal formula, leading to equation 3:

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

where $\Phi(h_0) = 0$ and $\Phi(h_n) = 1$ for construction. It will represent the percentage of the area between the *perfect equality line* and the *perfect inequality line*. This coefficient will be directly proportional to the inequality of the distribution. The interval of the Gini coefficient result will be from 0 to 1 (or 0% to 100%). A low value of Gini coefficient indicates a more equal distribution and a high value indicates a more unequal distribution. This amount seems to be also appropriated to quantify if the collaboration is equalitarian or not in social networks as we will describe in the next section.

4. GINI COEFFICIENT APPLIED ON SOCIAL NETWORKS ANALYSIS

Our proposal is to apply the Gini coefficient in the context of Social Networks Analysis. This application is possible because

the relational ties (edges) between actors (nodes) can be viewed as a distribution of amounts corresponding to the possible pairs of actors on the SN (each actor combined with all other your neighbors in the SN). For each pair of actors, the value associated is: 0 if there is no relationship between them and the normalized weight of the relationship between them (or 1 in unweighted SN) otherwise.

Thus, the Gini coefficient of this particular Social Network can be calculated. The distribution will be perfectly equalitarian when every pair of actor has the same weight. On the other hand, a completely nonequalitarian distribution will be found out 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.

5. CASE STUDY

Our proposal is to apply the Gini coefficient in the context of Social Networks Analysis, more precisely in a Co-Authorship Social Network. In this SN, the actors are the researchers and the relational ties are the research collaboration between pairs of researchers. This is a weighted network where the collaboration relationships among the researchers are modeled and evaluated. The weights are given by:

$$w_{ij} = \frac{n_{ij}}{n_i} \quad (4)$$

where n_{ij} denotes the number of common papers between the pair of the neighbors $\langle i, j \rangle$. Here $n_i = \sum_{j=1}^n n_{ij}$.

Clearly we are talking about non-symmetric weight since w_{ij} is different of w_{ji} when n_i is different of n_j . Here by construction n_i is the number of papers of author i .

The dataset used in the construction of studied networks and the analysis of Gini coefficient performed in this case study are presented in the next subsections.

5.1 Dataset Description

In this case study, we used a network composed by 27 researchers involved in a research project of the INWeb (MCT/CNPq Grant Number 573871/2008-6), the Brazilian National Institute of Science and Technology for the Web. This project begins at year 2008. All researchers are professors (including full, assistant, and associate professors) in a Brazilian major education institution (UFMG, UFRGS, UFAM and CEFET-MG) with graduate program in Computer Science.

Two networks of INWeb are constructed in this case study. The dataset used to networks construction was obtained at DBLP on August 03, 2010. These two networks are constructed with data of different time intervals. The time interval is related to the years of publication of the papers considered in the construction of the co-author network. The first time interval is the publications of the authors until the year 2007 and the network constructed with it is named as SN2007. The second time interval is the publications of the authors until the year 2010 and the network constructed with it is named as SN2010.

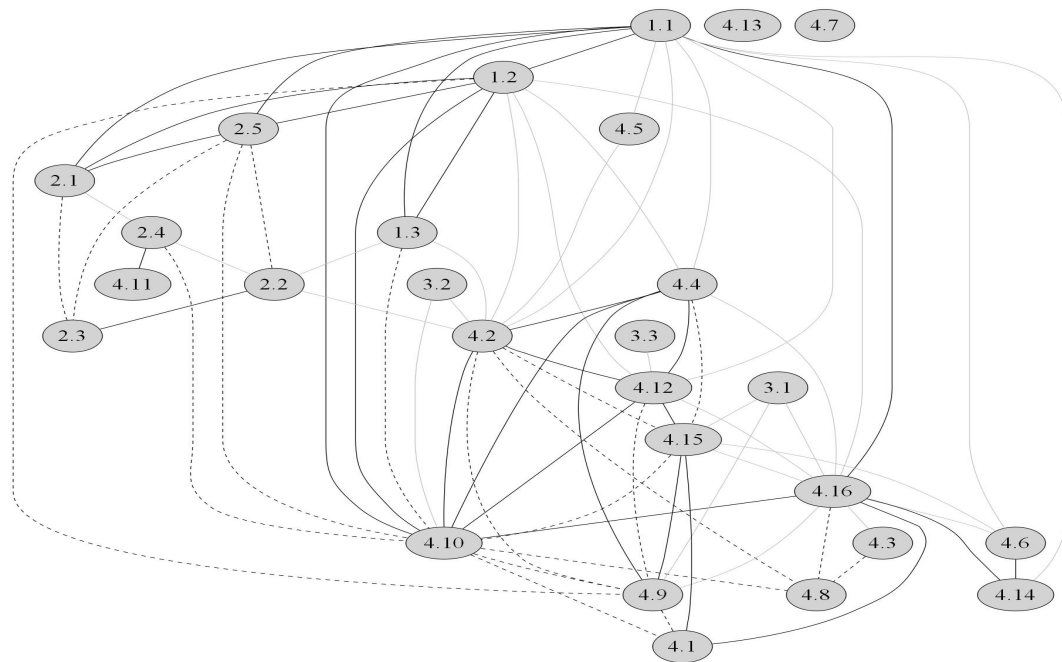


Figure 2. Comparative between Social Network of INWeb before the project beginning and the situation during the development of the project.

SN2007 reflects researchers' collaborations before the beginning of the INWeb project. SN2010 shows researchers' collaborations more than two years after the project beginning (before and during the development of the project). The case study aims to establish a comparative among the collaborations between researchers before and after the project implementation using the Gini Coefficient Analysis.

Figure 2 shows the modifications occurred in the researchers' collaborations considering a comparative between SN2007 and SN2010. It aims to establish a comparative between Social Network of INWeb before the project beginning and the situation during the development of the project. In Figure 2, the numbers in the nodes represent identifications of the researchers. The first digit (#Id) of these numbers identifies the institution affiliation of the researchers according to the Table 1. Researchers with the same first digit of identification are affiliated to the same institution. For a simplified representation, the authors that co-authored at least one paper are connected by only one line. The 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). The dashed lines represent the new connections between actors that occurred during the project until year 2010.

Table 1. The Institutions participating in the INWeb project

#Id	Institutions
1	UFAM
2	UFRGS
3	CEFET-MG
4	UFMG

5.2 Evaluations

Different evaluations are performed in the constructed networks of INWeb and they are presented and discussed as follows.

5.2.1 Global Analysis

The first analysis considers the distribution of weights of all possible relational ties between authors (including the weight equal to zero between researchers that 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 (distribution analyzed named SN2010_allPairs with value of Gini coefficient equal to 0.9327) than in SN2007 (distribution analyzed named SN2007_allPairs with value of Gini coefficient equal to 0.9471). These results are coherent and they are obtained because the SN2010 is more connected than SN2007. Moreover, it is possible to perceive that high values were obtained to both networks. These results indicate unequal distributions meaning that this collaboration network of INWeb is still very disconnected (in relation to one totally connected network) in both considered time intervals.

The INWeb research project has duration of four years and still stay in development. The results will be future evaluated, at the end of a four year period, the objective is that occurs an increasing in the cooperation patten 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.

Table 2. Gini Coefficient Values

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

5.2.2 Cooperative Analysis

The second analysis considers only the pairs of researchers that have some co-authored paper between them (weights nonzero) in one of the analyzed SNs (SN2007 or SN2010). This analysis of the Gini coefficient only among the cooperative pairs of researchers is important to evaluate the genuine level of collaboration between the pairs of researchers that effectively have relational ties modeled by the SN. The analysis considering only the pairs of researchers with nonzero weights in SN2007 aims to evaluate the pairs of researchers' that cooperated before the INWeb project beginning. Besides, the analysis considering only the pairs of researchers with nonzero weights in SN2010 aims to evaluate the pairs of researchers' that collaborated before and during the INWeb project development.

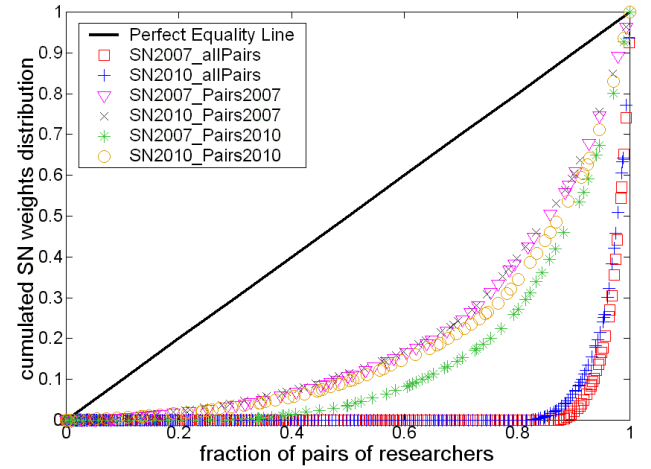
Cooperative Pairs until 2007. The results of analysis considering only the cooperative pairs of researchers until year 2007 (see Analysis 2.1 at Table 2) show that the value of Gini coefficient is lower in SN2010 (distribution analyzed named SN2010_Pairs2007 with value of Gini coefficient equal to 0.5735) than in SN2007 (distribution analyzed named SN2007_Pairs2007 with value of Gini coefficient equal to 0.5824). This difference reflects the improvement in the homogeneity of weight distributions occurred between the researchers that already collaborate before the beginning of the INWeb project. The collaborations between the researchers that are intensified, as can be observed in Figure 2, 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 (distribution analyzed named SN2010_Pairs2010 with value of Gini coefficient equal to 0.6160) than in SN2007 (distribution analyzed named SN2007_Pairs2007 with value of Gini coefficient equal to 0.7009). This difference reflects the improvement occurred in the collaborations of the researchers after the beginning of the INWeb project. This occurs because the pairs of authors considered in both SNs are equal. To this manner, the collaborations between the researchers are intensified, as can be observed in Figure 2, and the value of Gini coefficient identified this intensification of the connectivity. Moreover, it is possible to perceive that high values were obtained to both networks. These results indicate unequal distributions in both SNs. It is correct to both networks because: (i) in the SN2007 there are pairs of authors that are not connect and the distribution of weights is nonequalitarian, and (ii) in the SN2010, some already existent collaborations are intensified (increased weights) while new collaborations emerged (first

publications co-authored, probably low weights), then the distribution of weights is nonequalitarian in values.

5.2.3 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 3. In this figure, for the Lorenz curves construction, the cumulative percentage of pairs of INWeb researchers being analyzed is plotted on the x-axis; the cumulative percentage of the relationships weights of them, on the y-axis. Figure 3 presents the Lorenz curves of the distributions studied in the first and second analysis. The legend of the distributions is the same used in the Table 2.

**Figure 3. Lorenz Curves for the SNs distributions.**

In the second analysis (cooperative), only the pairs of researchers that have some relational tie between them were used. In this case, it is possible to perceive that lower areas between the Lorenz curves of the distributions and the perfect equality line were obtained when compared with the areas corresponding to the first analysis (global). These values indicate more equal distributions in second than in the first analysis, i.e. SNs more connected. These results are coherent by the set of data that were analyzed. This analysis of the Gini coefficient only among the cooperative pairs of researchers is important to evaluate the genuine level of collaboration between the researchers that have relational ties modeled by the social network.

In this case study, evaluations of Gini coefficient to analyze a real scientist social network were performed. Indeed, discussions about the obtained results were also presented. The case study presented satisfactory results. They show significant evidences of the validity and applicability of the Gini coefficient on the context of Social Networks.

6. CONCLUSIONS

This paper has as main contribution a new use of Gini Coefficient for analysis of scientist groups, more specifically in Social scientist Networks. Based on the contribution of scientists a network, a weighted graph is considered where the weight of each link is the number of common papers normalized for all contributions of the network and here the Gini coefficient measures the homogeneity level of the collaboration, i.e., if only a

few researchers keep a good level of collaboration or if all researchers of the network, indeed, are contributing for the group.

Pissard and Prieur [15] are performing similar techniques to quantify parameters in social networks. Although, they consider the problem where users in a social network are making upload of photos, in other hand, our work considers quantifying scientific collaboration in social networks.

As stated before, some of the metrics more commonly used for Social Networks Analysis originally does not consider the weights of relationships (if they exist) between the actors. These weights aim to measure the importance of the relational ties between actors and are important to be considered in a SNA. In our case study, we demonstrated the possibility of uses of the Gini Coefficient to analyze weighted Social Networks.

The initial results pointed out indices of the validity and utility of this approach. In further works, we intend to expand the study and to develop more evaluations. One possibility is to study if the Pareto principle can be verified in Social Networks distributions.

7. ACKNOWLEDGMENTS

G. R. Lopes, R. da Silva and J. P. M. de Oliveira would thank CNPq (Brazilian Research Council) and INWeb project (Brazilian National Institute of Science and Technology for the Web MCT/CNPq – Grant Number 573871/2008-6) for the partial financial support of this work.

8. REFERENCES

- [1] S. Wasserman, K. Faust, *Social Network Analysis: methods and applications*. Cambridge University Press, 1994.
- [2] D. Knoke, S. Yang, *Social Network Analysis*. 2.ed. [S.l.]: Sage Publications, Inc, 2007. 144p. (Quantitative Applications in the Social Sciences).
- [3] J. Barnes, "Class and Committees in a Norwegian Island Parish," *Human Relations*, [S.l.], n.7, p.39–58, 1954.
- [4] L. Freeman, "The impact of computer based communication on the social structure of an emerging scientific specialty," *Social Networks* 6, [S.l.], p.201–221, 1984.
- [5] Y. Ding, "Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks," *Journal of Informetrics* 5, n.1, p.187–203, 2011.
- [6] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, J. Guo, "Mining advisor-advisee relationships from research publication networks", *Proc. Of ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, Washington, DC, USA, p.203–212, 2010.
- [7] G. Menezes, N. Ziviani, A. Laender, V. Almeida, "A geographical analysis of knowledge production in computer science", *Proc. Of WWW*, p.1041–1050, 2009.
- [8] D. J. Watts, "Six Degrees: The Science of a Connected Age," W. W. Norton & Company, February 2003.
- [9] R. Recuero, "Redes Sociais na Internet: considerações iniciais," *E-Compós*, Brasília, v.2, 2005.
- [10] B. Hoser, A. Hotho, R. Jäschke, C. Schmitz, G. Stumme, "Semantic Network Analysis of Ontologies," *European Semantic Web Conference - ESWC 2006, Lecture Notes in Computer Science*, [S.l.], v.4011, p.514–529, 2006.
- [11] P. Marsden, "Egocentric and sociocentric measures of network centrality," *Social Networks*, 24, 407–422, 2002.
- [12] L. Freeman, "Centrality in social networks: I. Conceptual clarification," *Social Networks* 1, 215–239, 1979.
- [13] C. Gini. "Variabilità e mutabilità", 1912, Reprinted in *Memorie di metodologica statistica* (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi, 1955.
- [14] R. da Silva, J. Palazzo M. de Oliveira, J. V. de Lima, V. M. Orengo, "Statistics for Ranking Program Committees and Editorial Boards", <http://arxiv.org/abs/1002.1060>, 2010.
- [15] N. Pissard, P. Prieur, "Thematic vs. social networks in web 2.0 communities: A case study on Flickr groups", *Proc. Of Algotel Conference*, 2007.