Classical Statistical Inference for the Reliability of a Co-Authorship Network with Emphasis on Edges

Beatriz Barbero Brigantini
Sandra Cristina de Oliveira
Sergio Silva Braga Junior
Univ. Estadual Paulista – UNESP, Campus de Tupã
Street Domingos da Costa Lopes,780, Jd. Itaipu
Zip Code: 17602-496, Tupã, SP, Brazil

Abstract

Highly reliable research groups, i.e., with a strong collaborative framework of researchers, may contribute widely and intensely for the emergence and/or implementation of ideas, since they are responsible for most current research and also for the formation of numerous researchers. A research group may be considered a social network, which may be modeled by a graph. Researchers that make up this network may be interpreted as its nodes or actors, and the connections or links between these agents (represented by publications in common, i.e., co-authored papers) may be considered as its edges. In the literature, there are some ways to calculate the reliability of a network modeled by a graph G with k nodes and m edges. Current analysis measures the reliability of networks by taking into consideration unreliable edges and perfectly reliable nodes. Specifically, a statistical analysis based on classical inference to the network reliability has been proposed, obtaining the maximum likelihood estimators and confidence intervals for individual components (edges) and the network (probability of the research group to remain in activity at a given time t); the proposed methodology was applied to a research group of UNESP registered at CNPq; and measures of centrality of nodes were obtained to identify situations in which the insertion of an edge (connection between two researchers of the group) could significantly increase the reliability of a co-authoring network. Results show the feasibility of classical statistical inference coupled with the use of measures of centrality in the context of social network analysis.

Keywords: Social networks, research groups, theory of graphs, statistical inference.

1. Introduction

Reliability is the ability with which an item successfully performs a function under specific operational conditions. The term network reliability is bounded to the calculation of reliability of any general configuration of items (or components) when the reliability of each item is warranted.

Networks are physical, biological or social systems characterized by a huge set of well-defined items that interact dynamically. Physical networks comprise electricity and water distribution, transport, telecommunications, radio and TV, and others; social networks may be networks of personal and/or thematic relationships, communities, e-mails, blogs; biological networks may include food chains and disease transmission (LYRA & OLIVEIRA, 2011).

The maintenance of the functionality of a network requires information of its structure, functions and characteristics. Since a network´s structure may be represented by a graph, the Theory of Graphs is basic to determine the properties which refer to the network´s topological aspects.

Network reliability is the probability that a network remains functioning even though a flaw demands the removal of one or more subsets of the components (edges and/or nodes). Highly reliable networks are strong structures. Moreover, a network is more reliable than another if the probability of one network is disconnected is less than that of the other.

Everyone agrees that our planet has become more complex and that knowledge is more and more difficult to construct on an individual basis.

The stimulus to form research groups in universities and development organs proves this fact. The institutionalization of research groups in Brazil by the National Council for Scientific and Technological Development (CNPq) coupled to their dissemination and constant upgrading is a practice that foregrounds research in Brazil (MARAFON apud Miorin, 2008).Highly reliable research groups with a strong collaboration structure may contribute intensely towards the emergence and/or concretization of ideas. In fact, these groups perform most current research and are responsible for the formation of numberless researchers.

A research group is a social network and may be modeled by a graph. Researchers that form the network may be called its vertices or nodes and the connections and bonds between these nodes (for example, team publications) are the edges. Current analysis studies the reliability rate of networks when edges are unreliable or prone to flaws and the nodes are totally reliable. In other words, current research proposes (a) a statistical analysis based on the classical inference for a network´s reliability, with estimates of maximum likelihood and the respective confidence intervals for the reliability of edges (co-authorship bonds) individually and for the reliability of the network (the probability that the research group continues to function) at a given time *t*; (b) the development of an analysis for a special research group of the State University of São Paulo (UNESP) enrolled by the CNPq; (c) the provision of measures of centrality of nodes to identify situations in which the insertion of edges (or the co-authorship bond between two researchers of the team) may significantly increase the network´s reliability.

When scientific production is registered by the researchers and published on the CNPq Lattes Database (Lattes CV), the data may contain several types of imprecision (mistakes in the writing of names causing ambiguities and incorrect identification of authors; scientific articles under the name of one author but lacking under the name of the other co-authors; papers which unawares were not registered under any author, and others). Inference approach is, therefore, highly important for the reliability of co-author network.

2. Bibliographical Review

2.1 Social networks

Social networks are structures composed of people, organizations, territories or others, connected among themselves by one or several types of relationships (friendship, family, commercial etc.) through which information, knowledge, interests, values and aims (relationship, community, political, professionals networks) are shared. Social networks investigate the development of the team´s activity and indicate the group´s and the person´s efforts. Knowledge on the structure, function and traits of a social network are extremely relevant for its functionality. Since they may function at different levels, such as network of relationships, network of professionals, community network, political network etc, co-authorship networks are included in this context. In fact, they are made up of researchers and shared tasks. Networks are symmetrical in the sense that researcher A is a collaborator of researcher B at a given time *t* in the exact number of times that researcher B is a collaborator of A.

Since shared work or co-authorship saves time and financial and material resources, it is encouraged by research funding agencies in Brazil. These factors contribute towards the valorization of researchers that are capable of forming efficient and productive work teams (MAIA & CAREGNATO, 2008).

Regardless of certain particularities, co-authorship of products produced by scientific activities, with a special mention to scientific publications, indicates collaboration. Results on studies dealing with co-authorship reveal that collaboration among authors have increased significantly in all areas of knowledge and underscore the importance of co-authorship in the maintenance of research groups.

2.2 Basic concepts of the Theory of Graphs

The basic concepts of the Theory of Graphs were investigated by Boaventura Netto & Jurkiewicz (2009), Silva (2010) and Lyra & Oliveira (2011). Graph is a simple, abstract and intuitive notion which represents a sort of relationship between items. It is represented by a drawing with nodes or vertices which signify the items, bonded by lines, called edges, which denote the relationship.

The mathematic representation of a simple undirected graph is $G=(V,E)$, where V is the finite, not-empty set whose items are the nodes; E is a set of subsets of two items of V whose items are the edges. The set of nodes V has cardinality (number of items) $|V|=m$; the set of edges E has cardinality $|E|=k$; each edge is denoted by $\{v_i,v_j\}$, in which $v_i,v_j \in V$.

The degree of node v_i, denoted by $d(v_i)$, is the number of edges in the nodes. Two nodes are adjacent if an edge exists between them. A walk is a family of successively adjacent links. When the last link of the sequence is adjacent to the first, the walk is closed and called a circuit; contrastingly, it is open. A walk occurs when all edges of the graph are distinct. In this case, it is called a path.

When other nodes are reachable from any one of them, the graph is connected; otherwise, it is called unconnected. Edge connectivity, denoted by $\gamma(G)$, is the least number of edges whose removal transforms the graph into an unconnected G graph. The connectivity of the node, denoted $\kappa(G)$, is the least number of nodes whose removal (together with the edges bound to it) transforms the graph into an unconnected G graph. A G-generator sub-graph is a graph from G through the mere elimination of some of its edges (without making it unconnected).
Graph G with m nodes may de represented by a matrix, denoted by $A_{(G)}$ of the order m, called adjacency matrix of G in which entrance a_{ij} of the matrix is equal to 1 if v_i and v_j are adjacent; otherwise, it is equal to zero, for all $i, j = 1,2,...,m$.

Two graphs $G=(V_1,E_1)$ and $H=(V_2,E_2)$ are equal when $V_1=V_2$ and $E_1=E_2$. Isomorph graphs have the same structure; in other words, they have the same number of nodes and edges, albeit a different pattern.

2.3 Basic concepts for the Social Network Analysis

Since studies on social networks are interdisciplinary, several methodologies of analysis based on network structures are extant. Another methodology for the study of social networks is the Social Network Analysis (SNA) whose concepts are very similar to the Theory of Graphs, coupled to the mathematical language employed. Some concepts relevant to SNA, provided by Hayashi, Hayashi & Lima (2008) and Silva (2010) are given below.
Agents, items or nodes may be individual social units (people or firms) or collective social units (institutions, organizations, nations) where bonds establish relationships between the agents. Bonds may be classified as absent, weak and strong and are due to any type of liaison, such as consanguinity, friendship, professional and others. Relationship is a set of bonds with the same bonding criteria. In fact, relationships have two important features that condition the methods of data analysis available, or rather, direction and valorization. A relationship may be directional, when the agent is the transmitter and the other is the receiver (friendship; quotation etc), and non-directional, when the relationship is reciprocal (knowledge, co-authorship etc). In the case of valorization, relationship may be dichotomic (implies the presence or absence of a determined bond between two nodes) or valorized with discrete or continuous values (weight due to relationship; for instance, the number of scientific papers published in co-authorship by a certain number of researchers). The agent´s attributes are his/her individual characteristics, such as name, gender and age.

The tools most used in SNA comprise descriptive statistics (graphs, tables, distribution of frequencies, descriptive measurements and other); centrality measurements (degree of information, neighborhood and intermediation); cluster analysis (division of the network in subsets of agents constructed from bonds and the position they occupy).

2.4 Calculation of the network´s reliability

Let a network be modeled by a simple undirected graph $G=(V,E)$ with m nodes and k edges. So that the network functions (or in activity) at time t, every pair of nodes should be connected by at least one path. Let´s suppose that the nodes are reliable and only the edge tends to be faulty. Therefore, each edge i ($i = 1,2,...,k$) has a functioning probability (reliability of edge i) denoted by p_i. There are instances in which all the edges of a graph that models the network have the same functioning probability, simply denoted by p. Further, nodes are independent two by two. In other words, the failure of one does not imply the other´s failure. So that the reliability of a network (the probability graph G that models the network continues connected, even given the failure of one or more edges) may be calculated, the probability of each functioning stage of the network must first be determined:

$$\prod_{i \in E'} p_i \prod_{i \in (E \setminus E')} (1 - p_i) \qquad (1)$$

where E is the set of edges of graph G and E' is the set made up by the functioning edges of graph G. When the edges of graph G that models the network have the same functioning probability p, the network´s reliability is given as:

$$p_{R_G} = \sum_{i=m-1}^{k} S_i p^i (1-p)^{k-i} \quad (2)$$

where G is the graph that models the network with m nodes and k edges; S_i is the number of connected sub-graphs of G with i edges (KELMANS, 1966). When the edges of the graph that models the network have different functioning probabilities p_i, the reliability of the network p_{R_G} is calculated similarly as expression (2), or rather, when the connected sub-graphs of G with i edges are obtained, the probability of each functioning state of the network should be calculated and results added.

3. Methodological Procedures

3.1 Collection of data and the construction of the co-authorship network

The group called Research Center in Administration and Agribusiness (CEPEAGRO) of Applied Social Sciences area was selected so that a network of scientific co-authorship formed by researchers from a research group of UNESP could be constructed. If each researcher is represented by a node and two nodes are linked by one edge; if, and only if, the researchers have at least one publication in common, then the reliability of the research group during time *t* (represented by an undirected graph that models the co-authorship network among the researchers of this group) is the probability of the above-mentioned team to continue active during time *t*, even though one or more flaws (changes in the number of co-authorship´s relations) causes the removal of one or more subsets of the graph´s edges.

The following methodological procedures were undertaken in current analysis:

I. Survey of scientific production (articles in scientific journals, books, papers read in scientific events), published and listed by researchers on the Lattes database from the moment of their insertion in the research group. The set of data on the scientific production of each researcher required for the proposed analyses was composed of
   a. the number of publications of each researcher, attributed to the research group;
   b. the number of common (or co-authored) publications between research peers attributed to the research group;

II. Organization and systematization of collected data, coupled to the representation and analysis of the characteristics of collaboration (co-authorships in scientific publications) among the researchers under analysis, by a graph;

III. Calculation of three centrality measurements of nodes: 1) measurement of closeness; and, 2) measurement of degree of information. The above measurements identify situations in which the insertion of an edge or of a bond between two researchers of the group may significantly increase the network´s reliability.

3.2 Calculation of reliability of co-authorship network

As discussed above, a research group may be dealt with as a social network which may be modeled by a simple undirected graph $G=(V,E)$ with m nodes (researchers that compose the research group) and k edges (co-authorship bonds).Since nodes are utterly reliable and only the edges are prone to failure, the reliability of the network or the probability of the team remaining in activity during time *t*, even though one or more flaws remove one or more subsets of the graph´s edges, is provided by

- Edges have the same probability of functioning p: $p_{R_G} = \sum_{i=m-1}^{k} S_i p^i (1-p)^{k-i}$
- Edges have the possibility of functioning several p_i: the reliability of network p_{R_G} is calculated as in the previous expression; in other words, when the connected sub-graphs of G with i edges are obtained, the probability of each functioning state of the network must be calculated, and results added.

3.3 Statistical inference (or maximum likelihood)

Supposing a co-authorship network modeled by a simple undirected graph $G=(V,E)$ with m nodes and k edges, Y_i is a variable indicator for the functioning of i^{th} edge (or rather, the i^{th} relation of co-authorship; $i = 1,2,\ldots,k$; Y is the indicating variable for the functioning of the network. Therefore,

$$Y_i = \begin{cases} 1 & \text{, if } i \text{ th edge is functionin g} \\ 0 & \text{, if } i \text{ th edge is not functionin g} \end{cases}$$

$$Y = \begin{cases} 1 & \text{, if network is functionin g} \\ 0 & \text{, if network is not functionin g} \end{cases}$$

Let $p_i = P\{Y_i = 1\}; i = 1,2,\ldots,k$, the reliability of the i^{th} edge and $p_{R_G} = P\{Y = 1\}$ the network´s reliability. Therefore, under certain conditions of conditional independence of $Y_i's$ and also of $p_i's$, p_{R_G} depends on $h(p_1, p_2, \ldots, p_k)$, where *h* is any function of the reliabilities of the individual components $p_i's$ and depends on the network´s structure (series, parallel or any other general configuration) (BARLOW & PROSCHAN, 1981).

Further, $p_i's$ and p_{R_G} may be estimated by the maximum likelihood method, an estimate technique very common in statistical inference. The likelihood principle holds that, if the model is correctly identified, all information from the data on the parameters is contained in the likelihood function. The method, therefore, selects the estimators of the model´s parameters that maximize the probability to obtain a really observed sample.

In this case, the likelihood function for $\mathbf{p} = (p_1, p_2, ..., p_k)'$ is given by:

$$L(D \mid \mathbf{p}) = \prod_{i=1}^{k} \binom{n_i}{x_i} p_i^{x_i} (1 - p_i)^{n_i - x_i} \tag{3}$$

where $D = \{(n_i, x_i); i = 1,2,...,k\}$ is a set of data in which n_i is the sum of the publications of researcher *r* and of researcher *s* for the research group and x_i is the number of co-authored publications of researchers *r* and *s* for the research group, where $r = 1,2,\ldots,m$, $s = 1,2,\ldots,m$, with $r \neq s$, and *m* is the number of nodes of graph *G*, or rather, of researchers that form the network (OLIVEIRA & ACHCAR, 2000).

Let $l(\mathbf{p})$ be the natural logarithm of the likelihood function $L(D \mid \mathbf{p})$. So:

$$l(\mathbf{p}) = \ln L(D \mid \mathbf{p}) = \sum_{i=1}^{k} \left[\ln n_i! - \ln x_i! - \ln(n_i - x_i)! + x_i \ln p_i + (n_i - x_i)\ln(1 - p_i)\right] \tag{4}$$

Deriving expression (4) with regard to $p_i, i = 1,2,...,k$, the following likelihood equation is obtained:

$$\frac{\partial l(\mathbf{p})}{\partial p_i} = \frac{x_i}{p_i} - \frac{(n_i - x_i)}{(1 - p_i)} \tag{5}$$

Making equation (6) equal to zero, the maximum likelihood estimators (MLE) of $p_i, i = 1,2,...,k$ are obtained:

$$\hat{p}_i = \frac{x_i}{n_i} \tag{6}$$

Asymptotical distributions for estimators $\hat{p}_i, i = 1,2,...,k$ are expressed by $\hat{p}_i \overset{a}{\sim} Normal(p_i, I_{ii}^{-1}(\mathbf{p})), i = 1,2,...,k$, in which $I_{ii}^{-1}(\mathbf{p})$ is the i^{th} term of the diagonal of the inverse Fisher matrix information (CASELLA & BERGER, 2010). Let $I_{ii}(\mathbf{p})$ be Fisher matrix information, given by:

$$I_{ii}(\mathbf{p}) = \begin{pmatrix} -E\left(\frac{\partial^2 l(\mathbf{p})}{\partial p_1^2}\right) & -E\left(\frac{\partial^2 l(\mathbf{p})}{\partial p_1 \partial p_2}\right) & -E\left(\frac{\partial^2 l(\mathbf{p})}{\partial p_1 \partial p_3}\right) & \cdots & -E\left(\frac{\partial^2 l(\mathbf{p})}{\partial p_1 \partial p_k}\right) \\ -E\left(\frac{\partial^2 l(\mathbf{p})}{\partial p_2 \partial p_1}\right) & -E\left(\frac{\partial^2 l(\mathbf{p})}{\partial p_2^2}\right) & -E\left(\frac{\partial^2 l(\mathbf{p})}{\partial p_2 \partial p_3}\right) & \cdots & -E\left(\frac{\partial^2 l(\mathbf{p})}{\partial p_2 \partial p_k}\right) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -E\left(\frac{\partial^2 l(\mathbf{p})}{\partial p_k \partial p_1}\right) & -E\left(\frac{\partial^2 l(\mathbf{p})}{\partial p_k \partial p_2}\right) & -E\left(\frac{\partial^2 l(\mathbf{p})}{\partial p_k \partial p_3}\right) & \cdots & -E\left(\frac{\partial^2 l(\mathbf{p})}{\partial p_k^2}\right) \end{pmatrix} \tag{7}$$

where $l(\mathbf{p})$ is the natural logarithm of the likelihood function $L(D|\mathbf{p})$ defined in (4), whilst $\mathbf{p} = (p_1, p_2, ..., p_k)'$.

Once more deriving equation (5) with regard to p_i, derivates of second order are obtained which correspond to the diagonals of the matrix expressed by (7), given by:

$$\frac{\partial^2 l(\mathbf{p})}{\partial p_i^2} = -\frac{x_i}{p_i^2} - \frac{(n_i - x_i)}{(1-p_i)^2} \tag{8}$$

It should be noted that second order derivates (in the above-mentioned matrix) corresponding to $\frac{\partial^2 l(\mathbf{p})}{\partial p_i \partial p_j}, i \neq j, i = 1,...,k; j = 1,...,k$, are equal to zero. Intervals of confidence $100(1-\alpha)\%$ for $p_i, i = 1,2,...,k$ are given by:

$$p_i : \left[\hat{p}_i - z_{\alpha/2}\sqrt{\hat{I}_{ii}^{-1}(\mathbf{p})}]; \hat{p}_i + z_{\alpha/2}\sqrt{\hat{I}_{ii}^{-1}(\mathbf{p})} \right] \tag{9}$$

where $(1-\alpha)$ is the level of the confidence of the interval; point $z_{\alpha/2}$ is determined from a normal standardized distribution; $\hat{I}_{ii}^{-1}(\mathbf{p})$ is the i^{th} term of the diagonal of the inverse Fisher matrix information (CASELLA & BERGER, 2010).

Since the reliability of network p_{R_G} is a reliability function of the individual components $p_i, i = 1,2,...,k$, to obtain MLE of p_{R_G}, the invariance property of the estimators of maximum likelihood is employed. In other words, it is sufficient to take the estimators of maximum likelihood of p_i, expressed in (6), and substitute in p_{R_G}, obtaining $\hat{p}_{R_G} = h(\hat{p}_1, \hat{p}_2, ..., \hat{p}_k)$.

Let $g^*(\mathbf{p}) = \left[\frac{\partial p_{R_G}}{\partial p_1} \quad \frac{\partial p_{R_G}}{\partial p_2} \quad \frac{\partial p_{R_G}}{\partial p_3} \quad \cdots \quad \frac{\partial p_{R_G}}{\partial p_k} \right]$ be a vector of the order $1 \times k$ and let $\Sigma = I^{-1}(\mathbf{p})$ be the inverse Fisher matrix information expressed by (7). By the method Delta[1] (SEM, SINGER & PEDROSO-DE-LIMA, 2009), a $100(1-\alpha)\%$ confidence interval (asymptotic) for p_{R_G} is given by:

$$p_{R_G} : \left[\hat{p}_{R_G} \pm z_{\alpha/2}\sqrt{\hat{g}^*(\mathbf{p})\,\hat{\Sigma}\,\hat{g}^*(\mathbf{p})'} \right] \tag{10}$$

where $\hat{g}^*(\mathbf{p})$ and $\hat{\Sigma}$ are sample matrixes of $g^*(\mathbf{p})$ and Σ respectively.

3.4 Measurements of the nodes´ centrality

Centrality measurements are employed in SNA to verify the relevance of a node with regard to the others in a network. Through centrality measurements, nodes may be ordered according to their relative importance. Since power is a relation-derived characteristic, it may be associated to centrality measurements by showing power distribution within a network and the influence of nodes to dominate or influence other nodes.

Different centrality measurements are used for different types of relevance (position, flux, influence and others). Among extant measures, the following were employed (SILVA, 2010; LYRA & OLIVEIRA, 2011):

Closeness measurement relates total distance of a node to other nodes of the network, or rather, it indicates the access velocity of a node to another one in the network and shows the nodes that need improvement. Closeness measurement of node *i* (v_i) is calculated by

[1] If the distribution convergence of a parameter is known, then, by the Delta method, the distribution convergence of a function of this parameter is also known. The function should satisfy certain conditions such as being differentiable and continuous.

$$C_p(v_i) = \sum_{j=1}^{m} d_{v_i v_j} \quad (11)$$

where $d_{v_i v_j}$ represents the least distance between node i (v_i) and node j (v_j); m is the number of nodes in the network. The most central item of the network has the lowest rate of $C_p(v_i)$, or rather, the item that communicates with the highest speed with the other items of the network due to its structural position.

Information degree measurement gives relevance to a node due to the number of direct bonds that it establishes with the other nodes of the network. In other words, it evaluates direct interference (or immediate effect for time $t+1$) of a node in the other by the number of measurement unit paths originating from a node. The calculation of the information degrees measure of node i (v_i) is given by

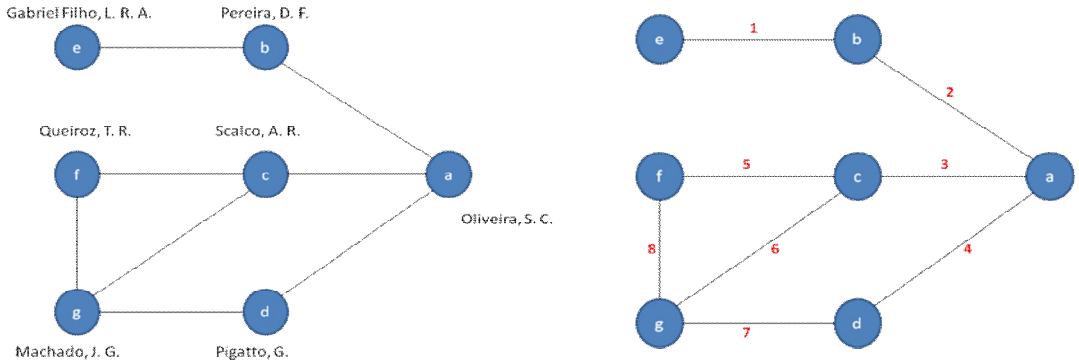
$$C_g(v_i) = d(v_i), 1 \le i \le m \quad (12)$$

where m is the number of nodes in the network.

4. Application, Results and Discussion

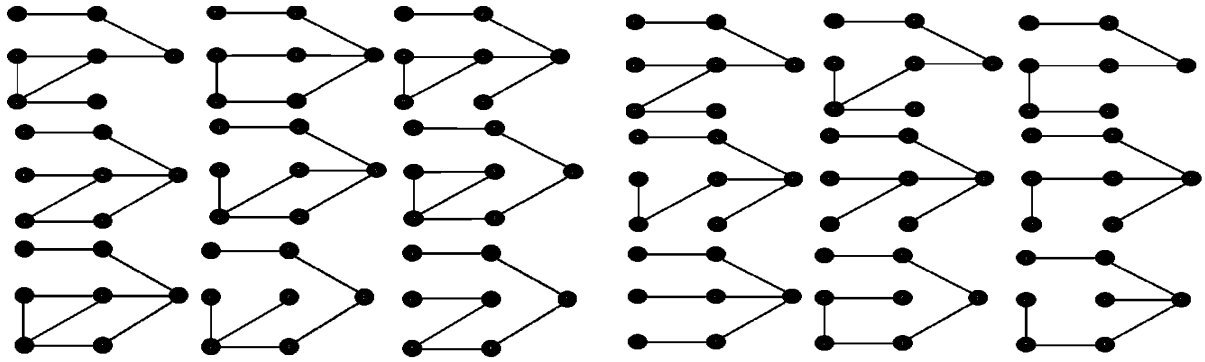
The graph modeling scientific co-authorship under analysis (CEPEAGRO) was automatically generated by script Lattes V7.02 according to the characteristics of collaboration between researchers of the group (co-authorship in scientific publications). Only articles in scientific journals, books and papers in scientific events were listed. They were filed and published by CEPEAGRO researchers at the Lattes Database from the date of inclusion up to August 2012 (time *t*). Researchers that quitted the group (at any moment since its establishment) were not taken into account. A network of scientific co-authorship modeled by undirected, simple, connected graph *G* was obtained, with $k = 8$ edges or co-authorship relations and $m = 7$ nodes or researchers, respectively.

Figure 1 – Graph *G* modeling the scientific co-authorship network.



According to Figure 2, eighteen connected sub-graphs may be formed from graph *G* of Figure 1. Given the configuration of graph *G*, it is impossible to form connected sub-graphs with five or less edges.

Figure 2 – Connected sub-graphs of *G* (Figure 1) with eight, seven and six edges.



Let us consider a (fictional) situation where all the edges of graph *G* have the same reliability $p_i, i = 1,2,...,8$ (denoted only by *p*), or rather, all co-authorship relations contribute equally for the group.

Therefore, the reliability of the network is given by $p_{R_G} = \sum_{i=6}^{8} S_i p^i (1-p)^{8-i}$, where S_i is the number of connected sub-graphs of *G* with *i* edges. Since $S_6 = 11$, $S_7 = 6$ and $S_8 = 1$, then:

$$p_{R_G} = S_6 p^6 (1-p)^{8-6} + S_7 p^7 (1-p)^{8-7} + S_8 p^8 (1-p)^{8-8} = 11p^6 (1-p)^2 + 6p^7 (1-p) + p^8 \quad (13)$$

Table1 shows the results of simulations for different *p* rates. The behavior of the reliability of the co-authorship network increases according to the reliability of each edge or co-authorship relation, or *p* rate. Due to the configuration of this group and the relation of existing co-authorship, the probability of flaw in the edge over 0.7 ($p < 0.3$) causes the network reliability (the probability that the group remains in activity during time *t*, on August 2012) to be close to zero.

Table 1: Reliability of scientific co-authorship network modeled by Graph *G* (edges with equal reliability), with different rates for *p*.

Values of *p*	Reliability of network p_{R_G}	Values of *p*	Reliability of network p_{R_G}
0,9	0,775904	0,4	0,022774
0,8	0,534774	0,3	0,004913
0,7	0,322358	0,2	0,000515
0,6	0,166095	0,1	0,00000946
0,5	0,070313		

Statistical inference

Let each edge *i*, $i = 1,2,...,8$, of graph *G* that models the co-authorship network has its reliability denoted by p_i. According to the eighteen connected sub-graphs from *G*, the generic reliability expression of the network is given by:

$$\begin{aligned} p_{R_G} &= p_1 p_2 p_3 p_4 p_5 p_6 p_7 p_8 + p_1 p_2 (1-p_3) p_4 p_5 p_6 p_7 p_8 + p_1 p_2 p_3 (1-p_4) p_5 p_6 p_7 p_8 + \\ &+ p_1 p_2 p_3 p_4 (1-p_5) p_6 p_7 p_8 + p_1 p_2 p_3 p_4 p_5 (1-p_6) p_7 p_8 + p_1 p_2 p_3 p_4 p_5 p_6 (1-p_7) p_8 + \\ &+ p_1 p_2 p_3 p_4 p_5 p_6 p_7 (1-p_8) + p_1 p_2 (1-p_3) p_4 (1-p_5) p_6 p_7 p_8 + p_1 p_2 (1-p_3) p_4 p_5 (1-p_6) p_7 p_8 + \\ &+ p_1 p_2 (1-p_3) p_4 p_5 p_6 p_7 (1-p_8) + p_1 p_2 p_3 (1-p_4)(1-p_5) p_6 p_7 p_8 + p_1 p_2 p_3 (1-p_4) p_5 (1-p_6) p_7 p_8 + \\ &+ p_1 p_2 p_3 (1-p_4) p_5 p_6 p_7 (1-p_8) + p_1 p_2 p_3 p_4 (1-p_5)(1-p_6) p_7 p_8 + p_1 p_2 p_3 p_4 (1-p_5) p_6 (1-p_7) p_8 + \\ &+ p_1 p_2 p_3 p_4 p_5 (1-p_6)(1-p_7) p_8 + p_1 p_2 p_3 p_4 p_5 (1-p_6) p_7 (1-p_8) + p_1 p_2 p_3 p_4 p_5 p_6 (1-p_7)(1-p_8) \end{aligned} \quad (14)$$

The reliability estimate process of each edge or relation of co-authorship *i* ($p_i, i = 1,2,...,8$) and the reliability of network p_{R_G} (research group in activity in August 2012) were undertaken by the maximum likelihood method. Consequently, according to information of scientific publications (papers in scientific journals, books and papers in events) of the research group CEPEAGRO obtained from script Lattes V7.02 and directly confirmed by the researchers, the set of data $D = \{(n_i, x_i); i = 1,2,...,8\}$ for the estimation process is given by:

Table 2: Set of data $D = \{(n_i, x_i); i = 1,2,...,8\}$ with regard to researchers of the research group CEPEAGRO.

Edge or co-authorship relation *i*	n_i	x_i
1	37	14
2	43	9
3	45	5
4	36	6
5	43	9
6	55	11
7	46	10
8	48	9

where n_i is the total (sum) of publications of researchers *r* and *s* for the research group; x_i is the number of co-authored publications of researchers *r*and *s* for the research group, in which $r = 1,2,\ldots,7$, $s = 1,2,\ldots,7$, with $r \neq s$. Likelihood function for $\mathbf{p} = (p_1, p_2, ..., p_8)'$ is given by $L(D \mid \mathbf{p}) = \prod_{i=1}^{8} \binom{n_i}{x_i} p_i^{x_i} (1 - p_i)^{n_i - x_i}$. Maximum Likelihood Estimators (MLE) of p_i are represented by $\hat{p}_i = \frac{x_i}{n_i}, i = 1,2,...,8$, and their respective $100(1-\alpha)\%$ confidence intervals (asymptotic) are expressed by $p_i : \left[\hat{p}_i \pm z_{\alpha/2} \sqrt{\hat{I}_{ii}^{-1}(\mathbf{p})} \right], i = 1,2,\ldots,8$. For a 95% confidence level, $z_{\alpha/2} = 1.96$ and $\hat{I}_{ii}^{-1}(\mathbf{p})$ are the i^{th} of the diagonal of inverse Fisher matrix information of expression (7), given by:

$$\hat{I}^{-1}(\mathbf{p}) = \begin{pmatrix} 157.31 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 259.83 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 259.20 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 455.63 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 259.83 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 343.75 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 270.38 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 315.08 \end{pmatrix}^{-1} \qquad (15)$$

According to expressions (14) and (15), MLE and the respective 95% confidence intervals for $p_i, i = 1,2,...,8$, are given in Table 3 below.

Table 3: **Maximum Likelihood Estimates (MLE) and 95% confidence intervals (CI) for reliabilities** $p_i, i = 1,2,...,8$.

	MLE	CI (95%)
p_1	0,3784	[0,2221 0,5347]
p_2	0,2093	[0,0877 0,3309]
p_3	0,1111	[0,0193 0,2029]
p_4	0,1667	[0,0449 0,2884]
p_5	0,2093	[0,0877 0,3309]
p_6	0,2000	[0,0943 0,3057]
p_7	0,2174	[0,0982 0,3366]
p_8	0,1875	[0,0771 0,2979]

Since the reliability of network p_{R_G} is a function of the reliability of individual components $p_i, i = 1,2,...,8$, the MLE of each $p_i, i = 1,2,...,8$ should be taken and substituted by p_{R_G} to obtain MLE of $\hat{p}_{R_G}$. In this case, the $100(1-\alpha)\%$ confidence interval for p_{R_G} is given by $p_{R_G} : \left[\hat{p}_{R_G} \pm z_{\alpha/2} \sqrt{\hat{g}^*(\mathbf{p}) \hat{\Sigma} \hat{g}^*(\mathbf{p})'} \right]$, where $\Sigma = I^{-1}(\mathbf{p})$ and $g^*(\mathbf{p}) = \left[\frac{\partial p_{R_G}}{\partial p_1} \quad \frac{\partial p_{R_G}}{\partial p_2} \quad \frac{\partial p_{R_G}}{\partial p_3} \quad \cdots \quad \frac{\partial p_{R_G}}{\partial p_8} \right]$. Thus,

$$\frac{\partial p_{R_G}}{\partial p_1} = p_2 p_3 p_4 p_5 p_6 p_7 p_8 + p_2(1-p_3) p_4 p_5 p_6 p_7 p_8 + p_2 p_3(1-p_4) p_5 p_6 p_7 p_8 +$$
$$+ p_2 p_3 p_4(1-p_5) p_6 p_7 p_8 + p_2 p_3 p_4 p_5(1-p_6) p_7 p_8 + p_2 p_3 p_4 p_5 p_6(1-p_7) p_8 +$$
$$+ p_2 p_3 p_4 p_5 p_6 p_7(1-p_8) + p_2(1-p_3) p_4(1-p_5) p_6 p_7 p_8 + p_2(1-p_3) p_4 p_5(1-p_6) p_7 p_8 +$$
$$+ p_2(1-p_3) p_4 p_5 p_6 p_7(1-p_8) + p_2 p_3(1-p_4)(1-p_5) p_6 p_7 p_8 + p_2 p_3(1-p_4) p_5(1-p_6) p_7 p_8 +$$
$$+ p_2 p_3(1-p_4) p_5 p_6 p_7(1-p_8) + p_2 p_3 p_4(1-p_5)(1-p_6) p_7 p_8 + p_2 p_3 p_4(1-p_5) p_6(1-p_7) p_8 +$$
$$+ p_2 p_3 p_4 p_5(1-p_6)(1-p_7) p_8 + p_2 p_3 p_4 p_5(1-p_6) p_7(1-p_8) + p_2 p_3 p_4 p_5 p_6(1-p_7)(1-p_8)$$

$$\frac{\partial p_{R_G}}{\partial p_2} = p_1 p_3 p_4 p_5 p_6 p_7 p_8 + p_1(1-p_3) p_4 p_5 p_6 p_7 p_8 + p_1 p_3(1-p_4) p_5 p_6 p_7 p_8 +$$
$$+ p_1 p_3 p_4(1-p_5) p_6 p_7 p_8 + p_1 p_3 p_4 p_5(1-p_6) p_7 p_8 + p_1 p_3 p_4 p_5 p_6(1-p_7) p_8 +$$
$$+ p_1 p_3 p_4 p_5 p_6 p_7(1-p_8) + p_1(1-p_3) p_4(1-p_5) p_6 p_7 p_8 + p_1(1-p_3) p_4 p_5(1-p_6) p_7 p_8 +$$
$$+ p_1(1-p_3) p_4 p_5 p_6 p_7(1-p_8) + p_1 p_3(1-p_4)(1-p_5) p_6 p_7 p_8 + p_1 p_3(1-p_4) p_5(1-p_6) p_7 p_8 +$$
$$+ p_1 p_3(1-p_4) p_5 p_6 p_7(1-p_8) + p_1 p_3 p_4(1-p_5)(1-p_6) p_7 p_8 + p_1 p_3 p_4(1-p_5) p_6(1-p_7) p_8 +$$
$$+ p_1 p_3 p_4 p_5(1-p_6)(1-p_7) p_8 + p_1 p_3 p_4 p_5(1-p_6) p_7(1-p_8) + p_1 p_3 p_4 p_5 p_6(1-p_7)(1-p_8)$$

$$\frac{\partial p_{R_G}}{\partial p_3} = p_1 p_2 p_4 p_5 p_6 p_7 p_8 - p_1 p_2 p_4 p_5 p_6 p_7 p_8 + p_1 p_2(1-p_4) p_5 p_6 p_7 p_8 +$$
$$+ p_1 p_2 p_4 p_5 p_6(1-p_7) p_8 + p_1 p_2(1-p_4)(1-p_5) p_6 p_7 p_8 + p_1 p_2(1-p_4) p_5(1-p_6) p_7 p_8 +$$
$$+ p_1 p_2(1-p_4) p_5 p_6 p_7(1-p_8) + p_1 p_2 p_4(1-p_5)(1-p_6) p_7 p_8 + p_1 p_2 p_4(1-p_5) p_6(1-p_7) p_8 +$$
$$+ p_1 p_2 p_4 p_5(1-p_6)(1-p_7) p_8 + p_1 p_2 p_4 p_5(1-p_6) p_7(1-p_8) + p_1 p_2 p_4 p_5 p_6(1-p_7)(1-p_8)$$

$$\frac{\partial p_{R_G}}{\partial p_4} = p_1 p_2 p_3 p_5 p_6 p_7 p_8 + p_1 p_2(1-p_3) p_5 p_6 p_7 p_8 - p_1 p_2 p_3 p_5 p_6 p_7 p_8 +$$
$$+ p_1 p_2 p_3 p_5 p_6(1-p_7) p_8 + p_1 p_2(1-p_3)(1-p_5) p_6 p_7 p_8 + p_1 p_2(1-p_3) p_5(1-p_6) p_7 p_8 +$$
$$+ p_1 p_2(1-p_3) p_5 p_6 p_7(1-p_8) + p_1 p_2 p_3(1-p_5)(1-p_6) p_7 p_8 + p_1 p_2 p_3(1-p_5) p_6(1-p_7) p_8 +$$
$$+ p_1 p_2 p_5(1-p_6)(1-p_7) p_8 + p_1 p_2 p_3 p_5(1-p_6) p_7(1-p_8) + p_1 p_2 p_3 p_5 p_6(1-p_7)(1-p_8)$$

$$\frac{\partial p_{R_G}}{\partial p_5} = p_1 p_2 p_3 p_4 p_6 p_7(1-p_8) + p_1 p_2(1-p_3) p_4(1-p_6) p_7 p_8 + p_1 p_2(1-p_3) p_4 p_6 p_7(1-p_8) +$$
$$+ p_1 p_2 p_3(1-p_4)(1-p_6) p_7 p_8 + p_1 p_2 p_3(1-p_4) p_6 p_7(1-p_8) + p_1 p_2 p_3 p_4(1-p_6)(1-p_7) p_8 +$$
$$+ p_1 p_2 p_3 p_4(1-p_6) p_7(1-p_8) + p_1 p_2 p_3 p_4 p_6(1-p_7)(1-p_8)$$

$$\frac{\partial p_{R_G}}{\partial p_6} = p_1 p_2(1-p_3) p_4(1-p_5) p_7 p_8 + p_1 p_2(1-p_3) p_4 p_5 p_7(1-p_8) + p_1 p_2 p_3(1-p_4)(1-p_5) p_7 p_8 +$$
$$+ p_1 p_2 p_3(1-p_4) p_5 p_7(1-p_8) + p_1 p_2 p_3 p_4(1-p_5)(1-p_7) p_8 + p_1 p_2 p_3 p_4 p_5(1-p_7)(1-p_8)$$

$$\frac{\partial p_{R_G}}{\partial p_7} = p_1 p_2(1-p_3) p_4 p_5 p_6 p_8 + p_1 p_2 p_3(1-p_4) p_5 p_6 p_8 + p_1 p_2(1-p_3) p_4(1-p_5) p_6 p_8 +$$
$$+ p_1 p_2(1-p_3) p_4 p_5(1-p_6) p_8 + p_1 p_2(1-p_3) p_4 p_5 p_6(1-p_8) + p_1 p_2 p_3(1-p_4)(1-p_5) p_6 p_8 +$$
$$+ p_1 p_2 p_3(1-p_4) p_5(1-p_6) p_8 + p_1 p_2 p_3(1-p_4) p_5 p_6(1-p_8) + p_1 p_2 p_3 p_4(1-p_5)(1-p_6) p_8 +$$
$$+ p_1 p_2 p_3 p_4 p_5(1-p_6)(1-p_8)$$

$$\frac{\partial p_{R_G}}{\partial p_8} = p_1 p_2 p_3 p_4(1-p_5) p_6 p_7 + p_1 p_2(1-p_3) p_4(1-p_5) p_6 p_7 + p_1 p_2(1-p_3) p_4 p_5(1-p_6) p_7 +$$
$$+ p_1 p_2 p_3(1-p_4)(1-p_5) p_6 p_7 + p_1 p_2 p_3(1-p_4) p_5(1-p_6) p_7 + p_1 p_2 p_3 p_4(1-p_5)(1-p_6) p_7 +$$
$$+ p_1 p_2 p_3 p_4(1-p_5) p_6(1-p_7) + p_1 p_2 p_3 p_4 p_5(1-p_6)(1-p_7)$$

When MLE of each $p_i, i = 1,2,...,8$ in the expression of the derivates $\frac{\partial p_{R_G}}{\partial p_1}$ $\frac{\partial p_{R_G}}{\partial p_2}$ $\frac{\partial p_{R_G}}{\partial p_3}$ … $\frac{\partial p_{R_G}}{\partial p_8}$ is substituted, $\hat{g}^*(\mathbf{p})$ is obtained. If the square root of $\hat{g}^*(\mathbf{p})\hat{\Sigma}\,\hat{g}^*(\mathbf{p})'$, with $\hat{\Sigma} = \hat{I}^{-1}(\mathbf{p})$ of the expression (15) is extracted and its result multiplied by $z_{\alpha/2} = 1.96$, then estimate error for the $100(1-\alpha)\%$ confidence interval for p_{R_G} is obtained. Table 4 shows maximum likelihood estimates of p_{R_G} and the respective 95% confidence interval.

Table 4: **Maximum Likelihood Estimates (MLE) and 95% confidence interval (CI) for the reliability of the network of scientific co-authorship p_{R_G} .**

	MLE	CI (95%)
p_{R_G}	0,0006587	[-0,000664 0,001982]

According to the configuration of the research group, number of researchers and co-authorship relations, coupled to data available on scientific production, the reliability of the network underscoring edges or co-authorship relations is very low. In fact, MLE of individual reliabilities of edges has respectively minimum and maximum rates of 11.11% and 37.84% (Table 3). Further, most MLEs of reliabilities varies closely to 20%. It should be underscored that edges or co-authorship relations 1 and 7 are respectively the most reliable, whereas co-authorship relations 3 and 4 have the lowest reliability rates in the network.

The maintenance and intensification of co-authorship relations among researchers of a research group are highly relevant for the maintenance of the group. As the reliability of such relationships increases, a rise in the reliability of the co-authorship network occurs.

Centrality measurements

Thirteen non-isomorph graphs may be generated from *G* within the possible options of insertions of a new edge in the co-authorship network modeled by graph *G* in Figure 1 (Table 5).

Table 5: **Non-isomorph graphs from *G* (Figure 1) with the insertion of a new edge.**

Graph	Co-authorship relation	Graph	Co-authorship relation	Graph	Co-authorship relation
*G*1	Pesquisadores ae e	*G*6	Pesquisadores b e f	*G*11	Pesquisadores d e f
*G*2	Pesquisadores a e f	*G*7	Pesquisadores b e g	*G*12	Pesquisadores eef
*G*3	Pesquisadores a e g	*G*8	Pesquisadores c e d	*G*13	Pesquisadores eeg
*G*4	Pesquisadores b e c	*G*9	Pesquisadores ce e		
*G*5	Pesquisadores b e d	*G*10	Pesquisadores de e		

Centrality measurements of closeness and information degree for nodes of graph *G* (Figure 1) were calculated, as Table 6 demonstrates.

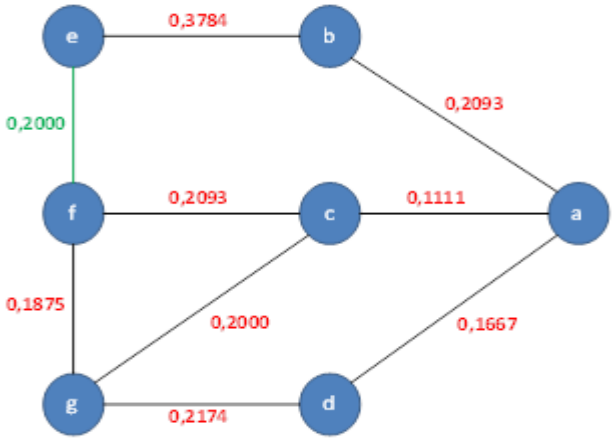
Table 6: **Centrality measurements of nodes of graph *G* of Figure 1**

Nodes or researchers of graph *G*	Closeness measurement	Information degree measurement
a	9	3
b	12	2
c	10	3
d	11	2
e	17	1
f	13	2
g	12	3

According to the above measurements, the most central nodes or researchers of graph *G* are respectively “a” and “c”, that is, the researchers with the highest speeds of access and with the greatest influence on the others. Although statistical inference shows that the co-authorship relation between researchers “a” and “c” are the least reliable, if they were removed somewhat from the graph, the scientific co-authorship network would be less connected and, consequently, its reliability would be severely compromised since some paths would not exist anymore.

The less central nodes of graph *G* are respectively "e" and "f". According to tests by some authors (LYRA & OLIVEIRA, 2011; OLIVEIRA, BRIGANTINI & UEHARA, 2013; SILVA, 2010), if it is aimed at making the research group more reliable with the insertion of a new edge or co-authorship relation, the centrality measurements indicate that the bond between researchers "e" and "f" may bring such improvement. When edge *i* = 9 with fictional reliability $p_9 = 0.20$ between nodes "e" and "f" (Figure 3) is fitted, and considering the other edges with reliabilities $p_i, i = 1,2,...,8$ equal to their respective maximum likelihood estimates in Table 3 and re-calculating expression (14), the network's reliability increases approximately 3.21 fold ($p_{R_G} = 0.002117$) when compared to the reliability of the network without the insertion of the above-mentioned edge($p_{R_G} = 0.000659$).

Figure 3: **Graph *G*12 of Table 5 (insertion of an edge between nodes "e" and "f") with edge reliability estimated $\hat{p}_i, i = 1,2,...,8$ and $p_9 = 0.20$.**



5. Final Considerations

Studies on the network reliability of scientific co-authorship identify which networks are reliable from different approaches (edges and/or nodes) according to the participation of researchers and the intensity of extant co-authorship relations.

Current investigation proposes a classical inference approach for the reliability of a co-authorship network with a specific focus on edges (co-authorship relations), or rather, taking into consideration perfectly reliable nodes (researchers). Further, centrality measurements of nodes were obtained that identified the situation in which the insertion of an edge between two researchers provided a significant increase in the reliability of the network or the research group in remaining active during a given time *t*.

The example provided showed that the calculation of reliability of a co-authorship network may be stressing when executed manually or by computer. The employment of centrality measurements may be considered a feasible alternative. However, some studies have shown that such measures may be an auxiliary alternative but not entirely reliable when investigating a network´s reliability increase (LYRA & OLIVEIRA, 2011; OLIVEIRA, BRIGANTINI & UEHARA, 2013; SILVA, 2010). Consequently, the use of other centrality measurements and the execution of simulations for more trust-worthy results are recommended besides the employment of these measurements.

Sponsoring Information:

The authors thank CNPq for financial support (proc. 406626/2012-0) and FAPESP for funding the scientific initiation scholarship (proc. 2012/01690-8).

6. References

Barlow, R. E. & Proschan, F. (1981).Statistical theory of reliability and life testing. New York: Holt, Rinehart and Winston.

Boaventura Netto, P. O. & Jurkiewicz, S. (2009).Grafos: Introdução e prática. São Paulo: Edgard Blucher.

Casella, G. & Berger, R. L. (2010).Inferência Estatística. São Paulo: Cengage Learning.

Hayashi, M. C. P. I., Hayashi, C. R. M. & Lima, M. Y. (2008). Análise de redes de coautoria na produção científica em educação especial. Revista Liincem, v. 4, n. 1, p. 84-103.

Kelmans, A. (1966). Connectivity of probabilistic networks. Automatic Telemekhania, v. 3, p. 98-116.

Lyra, T. F. & Oliveira, C. S. (2011). Um estudo sobre confiabilidade de redes e medidas de centralidade em uma rede de coautoria. Revista Eletronica Pesquisa Operacional para o Desenvolvimento, v. 3, n. 2, p. 160-172.

Maia, M. F. S. & Caregnato, S. E. (2008). Coautoria como indicador de redes de colaboração científica. Perspectivas em Ciência da Informação, v. 13, n. 2, p. 18-31.

Marafon, G. J. (2008). A importância dos grupos de pesquisa na formação de profissionais de Geografia Agrária: A experiência do NEGEF. Campo-Território: Revista de Geografia Agrária, v.3, n. 5, p. 284-290.

Morisette, J. T. & Khorram, S. (1998). Exact binomial confidence interval for proportions. Photogrammetric Engineering & Remote Sensing, v. 64, p. 281-283.

Oliveira, S. C. & Achcar, J. A. (2000). Confiabilidade de redes: Um enfoque Bayesiano. Revista de Matemática e Estatística, v. 18, p. 167-194.

Oliveira, S. C., Brigantini, B. B. & Uehara, J. K. (2013). Análise da confiabilidade de uma rede social fictícia com enfoque nas arestas e do uso de medidas de centralidade. In XX Simpósio de Engenharia de Produção.

Sen, P. K., Singer, J. M. & Pedroso-De-Lima, A. C. (2009). From finite sample to asymptotic methods in statistics. Cambridge: Cambridge University Press.

Silva, T. S. A. (2010). Um estudo de medidas de centralidade e confiabilidade de redes. (Master´s Dissertation in Technology) – Centro Federal de Educação Tecnológica Celso Suckow da Fonseca do Rio de Janeiro.