Social Networks 48 (2017) 110–115

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

A Bayesian approach for the reliability of scientific co-authorship networks with emphasis on nodes

Sandra Cristina Oliveira [a], Juliana Cobre [b,*], Taiane de Paula Ferreira [a]

[a] Universidade Estadual Paulista, 17602-496 Tupã, SP, Brazil
[b] Universidade de São Paulo, 13560-970 São Carlos, SP, Brazil

ARTICLE INFO

Article history:
Available online 3 September 2016

MSC:
91D30
90C35
62F15

Keywords:
Social networks
Graph theory
Research group
Bayesian inference
MCMC simulation methods

ABSTRACT

The co-authorship among members of a research group commonly can be represented by a (co-authorship) graph in which nodes represent the researchers that make up of this group and edges represent the connections between two agents (i.e., the co-authorship between these agents). Current study measures the reliability of networks by taking into consideration unreliable nodes (researchers) and perfectly reliable edges (co-authorship between two researchers). A Bayesian approach for the reliability of a network represented by the co-authorship among members of a real research group is proposed, obtaining Bayesian estimates and credibility intervals for the individual components (nodes or researchers) and the network. Weakly informative and non-informative prior distributions are assumed for those components and the posterior summaries are obtained by Monte Carlo-Markov Chain methods. The results show the relevance of an inferential approach for the reliability of scientific co-authorship network. The results also demonstrate that the contribution of each researcher is highly relevant for the maintenance of a research group. In addition, the Bayesian methodology was a feasible and easy computational implementation.

1. Introduction

The development of science usually is given through the joint work of several researchers that compose a network of co-authorship. This network of co-authorship is called highly reliable when is likely to continue producing science. If this probability decreases the reliability of the network also decreases. Knowing the reliability of a co-authorship network is the most direct way to identify the scientific collaboration within the academic milieu (Abbasi et al., 2010).

As the science is registered through some kind of publication, to analyze the reliability of a co-authorship network it is necessary to investigate the work of researchers that comprise it. For example, this analysis might consider the total number of publications of the involved researchers and the number of publications in common to two or more researchers. However, the records of scientific publications are prone to errors as misspellings or different forms of short names, different authors and same abbreviation (Barbastefano et al., 2013; Smalheiser and Torvik, 2009; Wang et al., 2012a,b, for example).

To the development of the science it is inevitable financial investment. Even when the financial resources are abundant they can not be misspent. Therefore it is essential to prioritize the more reliable research groups. In Brazil, the National Council for Scientific and Technological Development (CNPq), among others objectives, aims to foment scientific and technological research, to encourage and to recognize Brazilian researchers. Thus, knowing the flow of information and knowledge among its research is essential. The CNPq Lattes Platform (the Lattes Curriculum) is an information system that integrates curriculum database, research groups and institutions. The Lattes Curriculum contains a variety of information, has constant improvement and is nationally recognized by most development agencies, universities, and research institutes. As the inclusion of information is made by the researcher (commonly done in curriculum), the curriculums of Lattes Platform are subject to errors as some publications may not be included in the curriculums of some co-authors or appear incorrectly. Thus, the reliability of a research group or specifically a co-authorship network is not deterministic even in the research group listed at the CNPq.

The co-authorship among members of a research group is a social network since any structure in which the items are connected by some relationship can be considered as such. It is noteworthy that within a research group there are members interconnected

* Corresponding author. Fax: +55 16 3373 9650.
E-mail address: jucobre@icmc.usp.br (J. Cobre).

http://dx.doi.org/10.1016/j.socnet.2016.06.005
0378-8733/

by publications in common directly (co-authors) or indirectly (co-authors of co-authors). Generally, a network can be represented by a graph and its characteristics, as its reliability, are determined by its proprieties (Brigantini et al., 2014). So, in a (co-authorship) graph, relationship among these members can be represented by a graph in which the nodes represent the researchers and the edges represent the co-authorship. Newman (2004) cites several information that can be collected from a network and also various statistics each with its own purpose as clustering coefficient and density of a graph (see also Zare-Farashbandi et al., 2014; Arif, 2015), although not target network reliability. Kumar (2015) provides a recent review about these topics including other kind of analysis such as mathematical analysis.

The network reliability is given by the probability of this network to remain functioning even when one or more subsets of the components (edges and/or nodes) are removed (Barlow and Proschan, 1981). The greater the probability of this network continuing connected, the greater your reliability and then this network is called highly reliable (Brigantini et al., 2014). Consequently, the network reliability definition by itself allows the edges and/or nodes of the graph to be taken into consideration in its analysis.

Lyra and Oliveira (2011) obtain the reliability of the co-authorship network assuming values to the reliability of the researchers, what means that there is no inference process and no conclusive result since the analysis just calculates all the possibilities considering the assumptions. Brigantini et al. (2014) takes into account the publications between two researchers (edges) to obtain the reliability of the co-authorship network. In this case it is assumed that the reliability of the co-authorship network considers all the possibilities of dissolution of the partnership (without disconnecting the network). To estimate the reliability of the co-authorship network (Oliveira et al., 2014) consider all possible departing researchers modeling the reliability through the nodes of the graph. The results of Oliveira et al. (2014) are based on asymptotic assumptions which may not be valid when there are few observations in the data set and the estimation of the reliability of the co-authorship network depends on partial derivatives with respect to all the parameters that represent the functioning probabilities of the nodes, what may not be feasible.

Following Oliveira et al. (2014) current investigation considers a measure of the reliability of networks with emphasis on nodes or researchers, i.e., assuming unreliable nodes (researchers) and perfectly reliable edges (co-authorship between two researchers). The goal of this proposal is to develop a Bayesian inferential approach to the reliability of co-authorship network, obtaining estimates and credibility intervals for the individual components (nodes or researchers) and the co-authorship network. For the inferential method chosen, weakly informative and non-informative priors are assumed and the posterior summaries are obtained by Monte Carlo Markov Chain (MCMC) methods. The main advantages of the Bayesian inference are that it does not depend on asymptotic results and the network reliability is straightly (and easily) obtained from the MCMC results of the researchers' reliability. Another point that deserves to be mentioned is that it is possible to incorporate previous knowledge about the parameters when available. Finally, an analysis for a real research group registered at CNPq exemplifies the proposal of this paper.

2. Estimation of the reliability of a co-authorship network

In a co-authorship network each researcher is represented by a node and two nodes are linked by one edge when the represented researchers have at least one publication in common.

2.1. Calculation of the reliability of networks

In order for the network, modeled by a simple undirected graph G with k nodes and m edges, to be active at time t, every pair of nodes should be connected by at least one path. When one node fails and the graph continues connected it is considered a stage of the network. It is plausible to consider that the nodes are independent two by two, that is, if one node fails it does not imply that the other will fail. So the reliability of a network is the probability of a graph G remaining connected at time $t+1$ considering all the possibilities of stages of the network.

It is assumed that only the nodes can fail with probability $1-p_i$, $i=1,\ldots,k$, what means that the edges are completely reliable and the functioning probability of node i is p_i. Then the probability of each functioning stage V_l of the network p_{V_l} is given by

$$p_{V_l} = \prod_{i\in V'} p_i \prod_{i\in V\setminus V'} (1-p_i), \quad l=1,\ldots,L, \tag{1}$$

where V denotes the set of nodes of graph G, V' denotes the set of functioning nodes of graph G and L is the total number of possible stages of the network. Thus the reliability of the network is given by

$$p_{R_G} = \sum_{l=1}^{L} p_{V'_l}. \tag{2}$$

Barlow and Proschan (1981) highlighted that p_{R_G} depends directly on p_i, $i=1,\ldots,k$, and indirectly on the network's structure, since the configuration of the network (series, parallel or other) influences V' and L.

Goldschmidt et al. (1994) lead with a network in which the functioning probabilities of the nodes are the same, that is, $p_i=p$, for all $i=1,\ldots,k$. In this particular case the network's reliability given in (2) is rewritten as follows

$$p_{R_G} = \sum_{i=2}^{k} S_i p^i (1-p)^{k-i}, \tag{3}$$

where G is the graph that models the network with k nodes and m edges; S_i is the number of connected sub-graphs of G with i nodes.

2.2. Bayesian inference

The Bayesian approach for parameter inference derives from the combination of the likelihood function for $\boldsymbol{p}=(p_1,p_2,\ldots,p_k)$, given by (4), with a prior density that reflects previous knowledge on the distribution of parameters by Bayes rule $\pi(\boldsymbol{p}|D)\propto L(\boldsymbol{p}|D)\pi(\boldsymbol{p})$, where $\pi(\boldsymbol{p}|D)$ is a posterior distribution of parameters and reveals how these random variables are distributed after data had been observed (Box and Tiao, 1973).

According to Oliveira and Achcar (2000), the likelihood function for $\boldsymbol{p}=(p_1,p_2,\ldots,p_k)$

$$L(\boldsymbol{p}|D) = \prod_{i=1}^{k} \binom{n_i}{x_i} p_i^{x_i} (1-p_i)^{n_i-x_i}, \tag{4}$$

where $D=\{(n_i,x_i), i=1,2,\ldots,k\}$ is the set of observed data, with n_i and x_i denoting the total number of publications and the number of co-authored publications of the researcher i, respectively, at time t; and k is the number of researches of graph G.

Current analysis proposes two possible assumptions, (weakly) informative prior distributions and non-informative prior distribution as follows.

- Informative prior density with prior independence for parameters $\boldsymbol{p}=(p_1, p_2, \ldots, p_k)$. Consequently, the joint prior density for $\boldsymbol{p}$ is calculated by the product of Beta distributions given by (5)

$$\pi(\boldsymbol{p}) \propto \prod_{i=1}^{k} p_i^{a_i-1}(1-p_i)^{b_i-1}, \tag{5}$$

where $0<p_i<1$, a_i and b_i are known and positive constants (Oliveira and Achcar, 2000). When (4) and (5) are combined, the posterior distribution for $\boldsymbol{p}=(p_1, p_2, \ldots, p_k)$ is given by

$$\pi(\boldsymbol{p}|D) \propto \prod_{i=1}^{k} p_i^{x_i+a_i-1}(1-p_i)^{n_i-x_i+b_i-1} \tag{6}$$

or rather, $p_i|D \sim \text{Beta}(x_i+a_i\,;\, n_i-x_i+b_i)$, $i=1, 2, \ldots, k$.

- Jeffreys' non-informative prior density (when scanty or no previous information exists on the distribution of the parameters). Then, joint prior density for $\boldsymbol{p}$ is given by (Box and Tiao, 1973)

$$\pi(\boldsymbol{p}) \propto |I(\boldsymbol{p})|^{1/2}, \tag{7}$$

where $I(\boldsymbol{p})$ is the expected information matrix for $\boldsymbol{p}$ whose elements are given by $I_{ij}(\boldsymbol{p}) = -E\left(\partial^2 l(\boldsymbol{p})/\partial p_i p_j\right)$, $i,j=1,2,\ldots,k$, with $l(\boldsymbol{p})$ as the natural logarithm of the likelihood function $L(\boldsymbol{p}|D)$, or rather,

$$l(\boldsymbol{p}) = \ln L(\boldsymbol{p}|D) = \sum_{i=1}^{k}[\ln n_i! - \ln(x_i!) - \ln(n_i-x_i)! + x_i \ln p_i + (n_i-x_i)\ln(1-p_i)]. \tag{8}$$

Since $(\partial^2 l(\boldsymbol{p})/\partial p_i^2) = -(x_i/p_i^2) - ((n_i-x_i)/(1-p_i)^2)$, $E(X_i)=n_i p_i$ and $(\partial^2 l(\boldsymbol{p})/\partial p_i p_j)=0$, $i \neq j$, $i=1, 2, \ldots, k$, $j=1, 2, \ldots, k$, the expected information matrix $I(\boldsymbol{p})$ has diagonal elements equal to $(n_i/(p_i(1-p_i)))$, $i=1, 2, \ldots, k$, and the others elements equal to zero. Therefore, Jeffreys' prior density for parameters $\boldsymbol{p}=(p_1, p_2, \ldots, p_k)'$ is given by (Gelman et al., 1995)

$$\pi(\boldsymbol{p}) \propto \prod_{i=1}^{k} p_i^{(1/2)-1}(1-p_i)^{(1/2)-1}, \tag{9}$$

which corresponds to the Beta(1/2, 1/2). When (4) and (9) are combined, we have the posterior distribution for $(p_i, p_2, \ldots, p_k)'$ expressed as

$$\pi(\boldsymbol{p}|D) \propto \prod_{i=1}^{k} p_i^{x_i+(1/2)-1}(1-p_i)^{n_i-x_i+(1/2)-1}, \tag{10}$$

or rather, $p_i|D \sim \text{Beta}(x_i+1/2\,;\, n_i-x_i+1/2)$, $i=1, 2, \ldots, k$.

There are closed expression to the conditional posterior distributions of the parameters excepted to p_{R_G}. In this case some sampling algorithm is needed to get the results of p_{R_G}. Since conditional posterior distributions are standardized, Gibbs sampling algorithm (Gelfand and Smith, 1990) was used to produce samples and obtain inferences on the reliability of the individual components p_i and on the reliability of the network p_{R_G}. It is an iterative sampling scheme involving a Markov Chain whose transition kernels are formed by complete conditional distributions. Let $\pi(\boldsymbol{p}|D)$ be the probability density in which $\boldsymbol{p}=(p_1, p_2, \ldots, p_k)'$ and complete posterior conditional distributions $\pi(p_i|D, p_{-i})$, $i=1, 2, \ldots, k$, are available. Algorithm provides a generation alternative based on successive generations of complete conditionals, given by steps 1, 2 and 3 as follows.

1. Initialize the iteration counter of the chains at $j=1$ and choose the initial rates $\boldsymbol{p}^{(0)} = (p_1^{(0)}, p_2^{(0)}, \ldots, p_k^{(0)})$.
2. Obtain a new rate $\boldsymbol{p}^{(j)} = (p_1^{(j)}, p_2^{(j)}, \ldots, p_k^{(j)})$ from $\boldsymbol{p}^{(j-1)}$ by a successive generation of rates

$$p_1^{(j)} \sim (p_1|p_2^{(j-1)}, p_3^{(j-1)}, \ldots, p_k^{(j-1)})$$
$$p_2^{(j)} \sim (p_2|p_1^{(j)}, p_3^{(j-1)}, \ldots, p_k^{(j-1)})$$
$$\vdots$$
$$p_k^{(j)} \sim (p_k|p_1^{(j)}, p_2^{(j)}, \ldots, p_{k-1}^{(j)})$$

3. Update the counter from j to $j+1$ and repeat (2) and (3) until convergence.

As the number of iterations increases, the chain will approach equilibrium. Therefore, it may be assumed that convergence was provided within iteration whose distribution is arbitrarily close to the equilibrium distribution $\pi(\boldsymbol{p}|D)$.

Gelman and Rubin (1992) suggest a verification method of convergence based on techniques of analysis of variance. Convergence is accepted when variation between the chains is smaller than the variance within the chains, or rather, when the statistics $\sqrt{\hat{R}_i} = \sqrt{\left[((S-1)/S)+(((T+1)B)/TSW)\right](df/(df-2))}$, $i=1, 2, \ldots, k$, will be approximately equal to 1, where B represents the variance between the chains, W represents the variance within the chains, df are the degrees of freedom for the Student's t distribution, T is the total number of simulations (or chains), and S is the number of algorithm iterations. More details may be found in Gelman and Rubin (1992).

Let $p_i^{(r,s)}$, $i=1, 2, \ldots, k$, be the rates obtained for p_i at the rth replication and at the sth iteration. Then, a Monte Carlo (or Bayesian) estimate of p_i with regard to the function of quadratic loss, is expressed by

$$\hat{p}_i = \frac{1}{RS}\sum_{r=1}^{R}\sum_{s=1}^{S} p_i^{(r,s)} \tag{11}$$

where R is the total number of simulations (or chains) and S is the number of algorithm iterations. Therefore, a Monte Carlo estimate for the reliability of the network p_{R_G} is given by

$$\hat{p}_{R_G} = \frac{1}{RS}\sum_{r=1}^{R}\sum_{s=1}^{S} h(p_1^{(r,s)}, p_2^{(r,s)}, \ldots, p_k^{(r,s)}), \tag{12}$$

where $h(\cdot)$ is the function that represents the reliability of the network and it depends on the network structure (series, parallel or any other configuration) (Barlow and Proschan, 1981). Usually there is no general expression to express $h(\cdot)$ because it must consider all the possible subgraphs that do not disconnect the network. Expression (13) in Section 3 is an example of function $h(\cdot)$ which can not be reduced to a general expression.

3. Real data analysis

A research group entitled Research Center in Administration and Agribusiness (CEPEAGRO) and registered at CNPq was considered to illustrate the purpose of this study. The graph that represents the co-authorship among the members of this research group was automatically generated by *scriptLattes V7.02* and it was generated considering only papers in scientific journals, books and publication in scientific events included at the Lattes database up to August 2012 (time t). Only researchers who made up the network at time t were considered in the study.

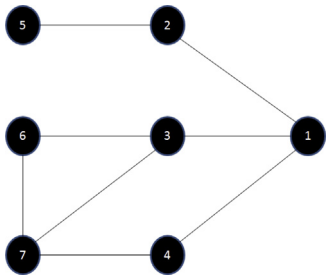

Fig. 1. Graph G modeling the co-authorship of the CEPEAGRO research group. *Source*: (Oliveira et al., 2014).

Fig. 1 shows the CEPEAGRO scientific co-authorship network modeled by undirected, simple, connected graph G with $m = 8$ edges or co-authorship relations and $k = 7$ nodes or researchers, that are given by (1) Oliveira; (2) Pereira; (3) Scalco; (4) Pigatto; (5) Gabriel Filho; (6) Queiroz and (7) Machado. These seven nodes, each one with reliability p_i, $i = 1, \ldots, 7$, and these eight edges lead to 43 connected sub-graphs, i.e., 43 connected sub-graphs can be formed from the graph G: a sub-graph with seven nodes, five sub-graphs with six nodes, nine sub-graphs with five nodes, eleven sub-graphs with four nodes, nine sub-graphs with three nodes and eight sub-graphs with two nodes. No sub-graph with a single node (researcher) is possible since the minimum number established by the CNPq for a research group requires two researchers. So the reliability of the CEPEAGRO network obtained from expression (2) is specifically given by expression (13).

Table 1
Data set $D = \{(n_i, x_i), i = 1, 2, \ldots, 7\}$ for the CEPEAGRO research group.

Node or researcher i	n_i	x_i	Node or researcher i	n_i	x_i
1	30	20	5	88	14
2	73	23	6	37	18
3	32	25	7	57	27
4	39	16			

$$
\begin{aligned}
p_{R_G} = {} & p_1p_2p_3p_4p_5p_6p_7 + p_1p_2(1-p_3)p_4p_5p_6p_7 + p_1p_2p_3(1-p_4)p_5p_6p_7 + p_1p_2p_3p_4(1-p_5)p_6p_7 + p_1p_2p_3p_4p_5(1-p_6)p_7 \\
& + p_1p_2p_3p_4p_5p_6(1-p_7) + p_1(1-p_2)p_3p_4(1-p_5)p_6p_7 + p_1p_2(1-p_3)p_4(1-p_5)p_6p_7 + p_1p_2(1-p_3)p_4p_5(1-p_6)p_7 \\
& + p_1p_2p_3(1-p_4)(1-p_5)p_6p_7 + p_1p_2p_3(1-p_4)p_5(1-p_6)p_7 + p_1p_2p_3(1-p_4)p_5p_6(1-p_7) + p_1p_2p_3p_4(1-p_5)(1-p_6)p_7 \\
& + p_1p_2p_3p_4(1-p_5)p_6(1-p_7) + p_1p_2p_3p_4p_5(1-p_6)(1-p_7) + (1-p_1)(1-p_2)p_3p_4(1-p_5)p_6p_7 + p_1(1-p_2)(1-p_3)p_4(1-p_5)p_6p_7 \\
& + p_1(1-p_2)p_3(1-p_4)(1-p_5)p_6p_7 + p_1(1-p_2)p_3p_4(1-p_5)(1-p_6)p_7 + p_1(1-p_2)p_3p_4(1-p_5)p_6(1-p_7) \\
& + p_1p_2(1-p_3)p_4(1-p_5)(1-p_6)p_7 + p_1p_2(1-p_3)p_4p_5(1-p_6)(1-p_7) + p_1p_2p_3(1-p_4)(1-p_5)(1-p_6)p_7 \\
& + p_1p_2p_3(1-p_4)(1-p_5)p_6(1-p_7) + p_1p_2p_3(1-p_4)p_5(1-p_6)(1-p_7) + p_1p_2p_3p_4(1-p_5)(1-p_6)(1-p_7) \\
& + (1-p_1)(1-p_2)(1-p_3)p_4(1-p_5)p_6p_7 + (1-p_1)(1-p_2)p_3(1-p_4)(1-p_5)p_6p_7 + (1-p_1)(1-p_2)p_3p_4(1-p_5)(1-p_6)p_7 \\
& + p_1(1-p_2)(1-p_3)p_4(1-p_5)(1-p_6)p_7 + p_1(1-p_2)p_3(1-p_4)(1-p_5)p_6(1-p_7) + p_1(1-p_2)p_3(1-p_4)(1-p_5)p_6(1-p_7) \\
& + p_1(1-p_2)p_3p_4(1-p_5)(1-p_6)(1-p_7) + p_1p_2(1-p_3)(1-p_4)p_5(1-p_6)(1-p_7) + p_1p_2(1-p_3)p_4(1-p_5)(1-p_6)(1-p_7) \\
& + (1-p_1)(1-p_2)(1-p_3)(1-p_4)(1-p_5)p_6p_7 + (1-p_1)(1-p_2)(1-p_3)p_4(1-p_5)(1-p_6)p_7 \\
& + (1-p_1)(1-p_2)p_3(1-p_4)(1-p_5)(1-p_6)p_7 + (1-p_1)(1-p_2)p_3(1-p_4)(1-p_5)p_6(1-p_7) \\
& + (1-p_1)p_2(1-p_3)(1-p_4)p_5(1-p_6)(1-p_7) + p_1(1-p_2)(1-p_3)p_4(1-p_5)(1-p_6)(1-p_7) \\
& + p_1(1-p_2)p_3(1-p_4)(1-p_5)(1-p_6)(1-p_7) + p_1p_2(1-p_3)(1-p_4)(1-p_5)(1-p_6)(1-p_7) \qquad (13)
\end{aligned}
$$

The estimation of reliability of each node or researcher i (p_i, $i = 1, 2, \ldots, 7$) and the reliability of the co-authorship network p_{R_G} was achieved by a Bayesian approach whose data set $D = \{(n_i, x_i), i = 1, 2, \ldots, 7\}$, in which n_i is the total number of publications of researcher i (either for the referred research group or for other aims) and x_i is the number of collaborators of researcher i for the research group under consideration, is presented in Table 1.

As non-informative prior distribution it will be considered the Jeffrey's non-informative prior distribution given in Section 2.2. Other natural choice would be the Beta(1, 1) distribution, that is the same of considering an Uniform(0, 1) distribution. The comparison between both was performed using the exact conditional posterior distributions and the differences between the posterior estimates were less than 0.3%. Under these circumstances efforts were saved assuming just the Beta(1/2, 1/2) distribution hereafter.

Practically, the rates of a_i and b_i in informative prior density may be attributed from any prior information on each researcher, for instance, the reliability of the researcher in other research group(s). The choices of a_i and b_i determine if the prior distribution is informative, strongly informative or weakly informative. In current analysis, each researcher was a priori considered moderately reliable. So that such a situation could be represented, the rates $a_i = 2.0$ and $b_i = 2.0$ were attributed from the measurements characteristics of Beta distribution, what leads to a weakly informative prior distribution.

Five chains of 2000 iterations each were produced for each case (weakly informative and non-informative priors) to obtain representative samples of posterior distributions and undertake inferences through simulation. The first 500 iterations of each chain were discarded to decrease the effect of initial conditions and took values spaced 15 to 15. Then, a sample of 500 data was produced. The algorithm was implemented using the MATLAB© software and the next graphics were constructed using software R (R Core Team, 2013).

Fig. 2 shows the proposed prior densities and their posterior densities of p_{R_G}. The posterior densities associated with the different prior distributions are similar, implying that the posterior distribution is not sensitive to the choice of the prior distributions. The following results corroborate this conclusion.

Table 2 shows Bayesian estimates, 95% credibility intervals (CI) and respective Gelman and Rubin (G–R) convergence indicators for p_i, $i = 1, 2, \ldots, 7$, taking into consideration non-informative prior density (NIPD) and weakly informative prior density (WIPD). Bayesian estimates of p_{R_G} and the respective 95% credibility intervals are provided by Table 3.

Table 2
Bayesian estimates, 95% credibility intervals and convergence indicators for reliabilities p_i, $i = 1, 2, \ldots, 7$.

		p_1	p_2	p_3	p_4	p_5	p_6	p_7
NIPD	Bayesian estimates	0.6613	0.3188	0.7692	0.4111	0.1626	0.4897	0.4758
	CI (95%)	[0.4752; 0.8234]	[0.2246; 0.4229]	[0.6281; 0.9093]	[0.2646; 0.5580]	[0.0962; 0.2421]	[0.3348; 0.6533]	[0.3660; 0.5983]
	G–R	0.9997	1.0029	1.0041	0.9965	1.0043	0.9990	1.0036
WIPD	Bayesian estimates	0.6416	0.3231	0.7507	0.4128	0.1757	0.4889	0.4759
	CI (95%)	[0.4701; 0.7977]	[0.2158; 0.4282]	[0.5997; 0.8807]	[0.2761; 0.5657]	[0.1026; 0.2662]	[0.3369; 0.6437]	[0.3524; 0.6028]
	G–R	1.0024	1.0026	1.0008	0.9977	1.0023	1.0095	1.0013

From these results, in the case of NIPD, it may be observed that, according to the configuration of the research group, number of researchers, and relationship between co-authors and data on the available scientific production, the estimated reliability for the co-authorship network was moderate (56.92%) and Bayesian estimates of researchers' reliability had a minimum rate of 16.26% and a maximum one of 76.92%.

Taking WIPD into consideration, estimated reliability for the co-authorship network was also moderate (56.71%) since Bayesian estimates for researchers' reliability had a minimum rate of 17.57% and a maximum one of 75.07%. The employment of the described WIPD did not present more accurate results in comparison with the NIPD, because the mentioned WIPD is weakly informative. Then it would be expected that the difference between the two prior densities chosen was small. However, in the case of relevant information on the reliability of each researcher by WIPD, improvement occurred in the precision of results since credibility intervals had less amplitude than those obtained by NIPD.

It also emphasizes that a sensitivity analysis was carried out and according to its results, when the prior information is scarce the posteriors means are not sensitive to the choice of the prior distributions. Of course that (highly) wrong prior information will misrepresent the results.

Regardless of prior (weakly informative and non-informative) density, it should be noted that researchers 3 and 1 are the most reliable, or rather, those with the highest relative contributions for the research group, respectively. On the other hand, researcher 5 is the least reliable. Finally, it has also been reported that the contribution of each researcher is highly relevant for the maintenance of the research group. As the collaboration of the researcher increases, his reliability also increases and, consequently, an increase in the reliability of the scientific co-authorship network may be detected.

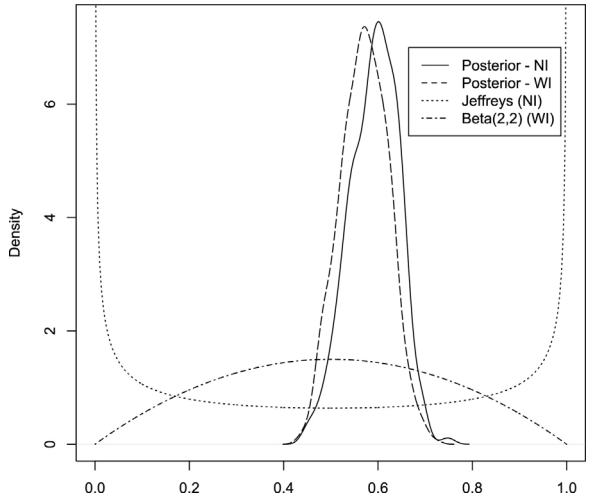

Fig. 2. Proposed prior densities (Jeffreys as non-informative – NI – and Beta(2, 2) as weakly informative – WI) and their posterior densities of p_{R_G}, (Posterior – NI and Posterior – WI).

Table 3
Bayesian estimates and 95% credibility intervals for the reliability of the scientific co-authorship network p_{R_G}.

		p_{R_G}
NIPD	Bayesian estimates	0.5692
	CI (95%)	[0.4883; 0.6813]
WIPD	Bayesian estimates	0.5671
	CI (95%)	[0.4703; 0.6600]

Table 4
Average of the amplitudes (amplitude) and coverage probability (cp) of the 95% credible intervals from 500 replications.

	NIPD		WIPD	
	Amplitude	cp	Amplitude	cp
p_1	0.3197	0.950	0.3149	0.964
p_2	0.2080	0.954	0.2068	0.950
p_3	0.2758	0.942	0.2820	0.948
p_4	0.2947	0.952	0.2858	0.940
p_5	0.1914	0.935	0.1991	0.930
p_6	0.3068	0.956	0.2973	0.942
p_7	0.2510	0.934	0.2459	0.960
p_{R_G}	0.1998	0.956	0.1905	0.953

4. Validation of the results

To analyze the behavior of the estimates it was conducted a simulation study in which were considered 500 replications of an artificial data set with similar characteristics to the real data set. With more details, to generate the 500 replications of the data set, they were considered groups of 7 researches following the network showed in Fig. 1, whose numbers of publications were fixed at $(n_1, \ldots, n_7) = (30, 73, 32, 39, 88, 37, 57)$. The researchers' reliability, p_i, $i = 1, \ldots, 7$, were generated from a $U(0, 1)$ and the number of collaborations of researcher i for the research group, x_i, $i = 1, \ldots, 7$, were generated from Binomial(n_i, p_i), $i = 1, \ldots, 7$.

The same estimation procedures considered in the real analysis were used here. Table 4 shows that the coverage probability of the 95% credible intervals for the estimates of p_i, $i = 1, \ldots, 7$, and p_{R_G} are around 95% what means that 95% of the credible intervals capture the true value of the parameters. Also according to Table 4 it is verified that in this scenario the averages of the amplitudes of the credible intervals are between 20% and 30%. Then it is possible to conclude that the obtained results are in agreement with what is expected.

5. Conclusion

Current investigation proposes a Bayesian inference approach for the reliability of a co-authorship network with a specific focus on nodes, that represent the researchers, and considering perfectly reliable edges, which represent the co-authorship relations. The methodology is feasible and it is easily implemented computationally. When there is prior information about the parameters of interested it usually is scarce, resulting in weakly informative prior

distributions. So, despite that this prior information could be incorporated in the estimation process, the results of this research show that there is no loss when are assumed non-informative prior distributions to the parameters in comparison with weakly informative prior distributions.

The advantage of a modeling focused on nodes is that it is possible, in a future research, to measure the importance of each researcher in the co-authorship network through the calculus of the network reliability conditioned on its absence. Although the approach considering both edges and nodes is desired there are two reasons for not considering it. First, because it can be intractable. And second, the extra information obtained with the addition of the edge to the approach considering just the nodes may be negligible.

It should be underscored that the inference approach proposed for the evaluation of reliability of co-authorship networks combined with social network analysis studies may generate relevant results for the maintenance of the functionality of research groups and their activities, which in turn contributes to create survival and competitiveness strategies for these groups.

References

Abbasi, A., Altmann, J., Hwang, J., 2010. Evaluating scholars based on their academic collaboration activities: two indices, the rc-index and the cc-index, for quantifying collaboration activities of researchers and scientific communities. Scientometrics 83 (1), 1–13.

Arif, T., 2015. The mathematics of social network analysis: metrics for academic social networks. Int. J. Comput. Appl. Technol. Res. 4 (12), 889–892.

Barbastefano, R.G., Souza, C., Costa, J.S., Teixeira, P.M., 2013. Impactos dos nomes nas propriedades de redes sociais: Um estudo em rede de coautoria sobre sustentabilidade. Perspectivas em Ciência da Informaç ao 18 (3), 78–95.

Barlow, R.E., Proschan, F., 1981. Statistical Theory of Reliability and Life Testing. Holt, Rinehart and Winston, New York.

Box, G.E.P., Tiao, G.C., 1973. Bayesian Inference in Statistical Analysis. Wiley Classics, New York.

Brigantini, B.B., Oliveira, S.C., Braga Junior, S.S., 2014. Classical statistical inference for the reliability of co-authorship network with emphasis on edges. Am. Int. J. Contemp. Res. 4 (3), 19–31.

Gelfand, A.E., Smith, A.F.M., 1990. Sampling-based approaches to calculating marginal densities. J. Am. Stat. Assoc. 85 (410), 398–409.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7 (4), 457–472.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis, 2nd edition. Chapman & Hall, London.

Goldschmidt, O., Jaillet, P., Lasota, R., 1994. On reliability of graphs with node failures. Networks 24 (4), 251–259.

Kumar, S., 2015. Co-authorship networks: a review of the literature. Aslib J. Inform. Manage. 67 (1), 55–73.

Lyra, T.F., Oliveira, C.S., 2011. Um estudo sobre confiabilidade de redes e medidas de centralidade em uma rede de co-autoria. Pesquisa Operacional para o Desenvolvimento 3 (2), 160–172.

Newman, M., 2004. Who is the best connected scientist? A study of scientific coauthorship networks. Complex Netw., 337–370.

Oliveira, S.C., Achcar, J.A., 2000. Confiabilidade de redes: Um enfoque bayesiano. Revista de Matemtica e Estatstica 18, 167–194.

Oliveira, S.C., Ferreira, T.P., Brigantini, B.B., Uehara, J.K., 2014. Classical statistical inference for the reliability of co-authorship network with emphasis in nodes. Perspectivas em Ciência da Informaç ao 19 (4), 202–225.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria http://www.R-project.org/.

Smalheiser, N.R., Torvik, V.I., 2009. Author name disambiguation. Annu. Rev. Inform. Sci. Technol. 43 (1), 1–43.

Wang, D.J., Shi, X., McFarland, D.A., Leskovec, J., 2012a. Measurement error in network data: a re-classification. Social Netw. 34 (4), 396–409.

Wang, J., Hicks, D., Melkers, J., Xiao, F., Pinheiro, D., 2012b. A boosted-trees method for name disambiguation. Scientometrics 93 (2), 391–411.

Zare-Farashbandi, F., Geraei, E., Siamaki, S., 2014. Study of co-authorship network of papers. J. Res. Med. Sci. 19 (1), 41–46.