Brazilian Journal of Probability and Statistics
2018, Vol. 32, No. 4, 755–782
https://doi.org/10.1214/17-BJPS362

Brazilian network of PhDs working with probability and statistics

Luciano Digiampietri[a], Leandro Rêgo[b], Filipe Costa de Souza[c], Raydonal Ospina[c] and Jesús Mena-Chalco[d]

[a]*Universidade de São Paulo*
[b]*Universidade Federal do Ceará*
[c]*Universidade Federal de Pernambuco*
[d]*Universidade Federal do ABC*

Abstract. Statistical and probabilistic reasoning enlightens our judgments about uncertainty and the chance or beliefs on the occurrence of random events in everyday life. Therefore, there are scientists working with Probability and Statistics in various fields of knowledge, what favors the formation of scientific network collaborations of researchers with different backgrounds. Here, we propose to describe the Brazilian PhDs who work with probability and statistics. In particular, we analyze national and states collaboration networks of such researchers by calculating different metrics. We show that there is a greater concentration of nodes in and around the cites which host Probability and Statistics graduate programs. Moreover, the states that host P&S Doctoral programs are the most central. We also observe a disparity in the size of the states networks. The clustering coefficient of the national network suggests that this network and regional differences especially with respect to states from South-east and North is not cohesive and, probably, it is in a maturing stage.

1 Introduction

Traditionally, academic collaboration is represented via co-authorship network (Glänzel and Schubert (2005), Yoshikane and Kageura (2004), Newman (2001), Newman and Girvan (2004), Neal (2014) and Stefano et al. (2013)). The use of co-authorship is especially useful for being a well-defined relationship, because it is easy to obtain data, and it is possible to replicate or update such studies. However, Katz and Martin (1997) argue that academic collaboration may be something much broader than co-authorship of scientific papers, including advisor-advisee relationships, and partnership in projects, classes, etc. Moreover, Melin and Persson (1996) warn that academic collaboration can also produce other products such as patents, or generate (in less than 5% of the cases, as estimated by the authors) no tangible product at all.

Key words and phrases. Academic collaboration, CNPq's productivity research fellows, probability and statistics, social network analysis.
Received April 2016; accepted April 2017.

Moreover, many performance studies used co-authorship metrics to explain academic performance using metrics such as number of articles written in English Yousefi-Nooraie et al. (2008) g-index Abbasi, Altmann and Hossain (2011), h-index Cimenler, Reeves and Skvoretz (2014) and research funds Bellotti (2012). These works traditionally show that nodes position and/or types of relationship play an important role in academic productivity.

In this paper, we analyzed the collaboration network among PhDs working with Probability and Statistics (P&S) in Brazil. Some reasons motivate us to perform this work, for example, the majority of the studies about the relation between network and performance metrics are based on co-authorship. Here, we also analyze the academic social network of Brazilian states, where the ties are not limited to co-authorship, including participation in projects and advisor-advisee relationship. On the other hand, to our knowledge, there is no social network study devoted to analyze Brazilian researchers working in the P&S field;

For the network design, the information contained in the Lattes platform (http://lattes.cnpq.br) was considered, and the relationships analyzed include: coauthorship, participation in a research project and the advisor-advisee relationship. These relationships were examined along the last 35 years (from 1980 to 2014). Different metrics were calculated for national and states collaboration networks of such researchers. We show that cities hosting graduate programs in P&S aggregate the majority of PhDs in these fields and states networks are heterogeneous in their sizes. Finally, the analysis of clustering coefficient of the national network indicates an immature stage of this network.

The remaining of this paper is organized as follows: Section 2 summarizes related works regarding to: the P&S area in Brazil, social network background and academic networks. Section 3 contains a brief description of Lattes platform from where the information of collaboration among Brazilian PhDs working with P&S was collected. Section 4 describes the methodology used for the development of this work. In Section 5, the results are presented and discussed. Section 6 contains the conclusions and directions for future work.

2 Related work

There are different types of works related to this one, for example: about the domain, the theoretical background and related to the methodology. The first, discussed in Section 2.1, is composed of works which analyze the P&S area in Brazil. The second, Section 2.2 present the main concepts of social networks analysis; and the third contains works that use academic curricula in order to assess academic social networks.

2.1 Probability and statistics in Brazil

Over the centuries, the probabilistic reasoning, and statistical and experimental methods are walking hand-by-hand with other scientific fields. It is almost impossible to imagine how the society could have evolved without that knowledge. From the medical industry to telecommunications, we are all surrounded by P&S applied knowledge. Although many theoretical studies are dealing to improve P&S methods, there are a much larger body of works applying it, and that is the main characteristic which makes P&S area so especial, that is, the massive interaction with other fields.

Therefore, it is plausible to imagine that there is a rich collaboration environment among those working with P&S. However, we still have little bibliometric information about this community, especially in Brazil. In this paper, we explore the scholarly networks of PhDs working with P&S in Brazil, considering the academic relationships from 1980 to 2014. To better understand the history of probability and statistics in Brazil, we recommend the following readings: Senra (2008, 2009) and Ara and Louzada (2012).

According to the Brazilian Ministry of Education website (e-MEC[1]) there are 81 undergraduate courses in statistics (at University of São Paulo, there is also a BA in Applied and Computational Mathematics with an emphasis on economic statistics) in Brazil. In this context, it is noteworthy that, as stated by Ara and Louzada (2012), knowledge about statistics permeate virtually all undergraduate courses in Brazil, and for being an evidence-based science, assists in the scientific development of different areas.

The P&S Graduate Courses are evaluated by the Mathematics, Probability and Statistics Committee from CAPES.[2] The result of the last assessment showed that there are nine Statistics Graduate Courses (one of them is a Mathematics and Statistics Graduate Course, and other is an Applied Mathematics and Statistics Graduate Course). From these nine courses, six have a PhD Program.[3] Moreover, from 2010 to 2012 these programs graduate 79 PhDs, with the largest contribution being made by USP (University of São Paulo) with 36 defenses. Regarding the CNPq research productivity fellowship (a fellowship targeted at researchers who stand out among their peers, enhancing their scientific production according to normative criteria established by CNPq), there are currently 70 fellows in payroll, 38 level 2, 8 level 1D, 4 level 1C, 15 level 1B and 5 level 1A,[4] being 1A the highest level and the 2 the lowest one.

[1] http://emec.mec.gov.br/ (accessed on 11/11/2015).
[2] Brazilian Coordination for the Improvement of Higher Education.
[3] http://www.avaliacaotrienal2013.capes.gov.br (accessed on 11/11/2015).
[4] http://plsql1.cnpq.br/divulg/ (accessed on 01/14/2015).

Ara and Louzada (2012), through a sample survey, describe the profile of professors of undergraduate Statistics courses at Brazilian Public Universities, especially regarding academic education. The authors found that most professors (63%) are not statisticians. Besides, among those with master's degree, only 31% has a master degree in statistics, and among the PhDs, only 20% have the PhD degree in statistics. The North and Northeast regions have the highest percentage of professors with an undergraduate degree in statistics (60%), on the other hand, in the South, only 12% have an undergraduate degree in statistics. The Southern region has the highest percentage of professors with Doctoral degrees (83%) and the North has the lowest percentage (38%). But among doctors, the scenario is reversed: the South has the lowest percentage of professors with a PhD degree in statistics (15%), while the northern region has the highest percentage (33%).

2.2 Social network background

According to Easley and Kleinberg (2010), a network or a graph is way to represent relationships among a collection of elements, named nodes or vertices. Formally, a network is a pair (N, M), where $N = \{1, 2, \ldots, n\}$ is a collection of elements or simply the finite set of nodes, and M is a $n \times n$ matrix, where m_{ij} represents the relationship between node i and node j. If nodes i and j in N are related, then we say that there is a link (or an arrow, or an edge, or a tie) between them. Depending on some characteristics of M, the network could be classified in different manners. When, for all i and j in N, $m_{ij} = m_{ji}$ the graph is said to be direct, otherwise, it is called undirected. Undirected graph happens when the relationship is not reciprocal, for example, an advisor-advised relationship. Moreover, when all values in M are taken from $\{0, 1\}$ the graph is said to be unweighted, where $m_{ij} = 1$ express that nodes i and j are related and $m_{ij} = 0$ indicates the absence of relationship. On the other hand, if the values of M could assume more than two values (expressing the intensity of the relationship) then the network is said to be weighted (Jackson (2008)).

As stated by Digiampietri and da Silva (2011), the characterization of a network to be direct (or not) or to be unweighted (or not) depends on the type of relationships analyzed. In academic networks, the relationships among nodes traditionally represent co-authorship or others tips of academic interaction such as advisor-advised relationship, partnership in projects etc., so the values in M are always non-negative integers.

De Stefano, Giordano and Vitale (2011) report that some academic connections are clearly direct or weighted, most of academic networks are treated as undirected and unweighted, especially when they also involve co-authorship or multiple relationships. This is justified because for many researches, the main goal is to identify (and understand) the relationship between academics, institutions and countries. In this context, to transform a weighted graph to its unweighted version, one shall simply to set all values in M that are greater than zero to one; and to transform

a direct graph to its undirected form, one shall set $m_{ij} = m_{ji} = 1$ every time that one pair of nodes i and j have $m_{ij} \neq m_{ji}$.

So, in what follows, we will discuss some concepts and metrics related to unweighted and undirect networks as could also be seen in Jackson (2008) and Mena-Chalco, Digiampietri and Cesar (2012):

- Total number of links: a link between two nodes indicates that they maintained a relationship in the network. Therefore, the total number of links indicates the total number of connections in the network during the analyzed period.
- Connected network: a network is connected if all its nodes can reach one another by a sequence of ties.
- Component of a network: a component of a network (N, M) is a connected subnetwork (N', M') where N' is a nonempty subset of N and M' is a submatrix of M such that if $i \in N'$ and $m_{ij} = 1$ in M, than $j \in N'$ and $m_{ij} = 1$ is preserved in M'.
- Size of component: is the total number of nodes in a given component. The biggest component in the network is called the giant component.
- Maximum clique size: correspond to the maximum subset of vertices in which everybody is related with each other.
- Degree of a node (or Degree Centrality): is the number of ties involving a given node. Nodes with degree equal to zero are called isolated.
- Average degree: is the sum of the degree of each node in the network divided by the total number of nodes.
- E-I index: is a segregation metric (Bojanowski and Corten (2014)), proposed by Krackhardt and Stern (1988), to evaluate the relationship between external and internal links in a network. By simplicity, suppose N was partitioned into two non-empty disjoint groups (one called internal group (IG) and the other called the external group (EG)). Let EL be the total number of links between nodes from IG and the nodes from EG (i.e., we only count a tie if one node is from IG and the other from EG); let IL be the total number of links only between nodes from IG and, finally, let $T = EL + IL$. Then the E-I index (EI) is formulated as $EI = (EL - IL)/T$. This index ranges from -1 (expressing that all links are internals) to $+1$ (expressing that all links are externals).
- Density: indicates the ratio between the number of edges in the network and the maximum number of possible edges, that is, it indicates how close the network is to be complete. Density equals to zero means that there are no edges in the graph, on the other hand, density equal to one means that all nodes are connected to each other.
- Diameter: is the maximum distance (or path) between any two nodes in the network. However, if the network is disconnected, then, the diameter of the network will be the biggest one among the diameters of each network component. So, the diameter could vary from one (in the best case scenario) to $\#N - 1$ (in the worst case scenario).

- Closeness Centrality: is the inverse of the average distance between a given node and all others nodes in its component.
- Betweenness centrality: is the average proportion of short paths that a given node lied on. Therefore, this metric indicates how important a node is to link other vertices.
- Eigenvector centrality: express the importance of a node in the network based on the importance of its neighbors; that is, as stated by Bonacich and Lloyd (2001), this metric is relevant when nodes' status is determined by their neighbors.
- Centralization: for each centrality metric (e.g., degree, betweenness, closeness and eigenvector), it indicates (based in a specific centrality measure) how central is the most central node in the network. These metrics are based on the sum of the differences between the most central vertex and all other vertices, divided by the theoretical maximum sum of differences. For more details, see Freeman (1978).
- Cluster coefficient: express the proportion of the vertices of a given node who also have a link between them (Latapy, Magnien and Vecchio (2008)). The average cluster coefficient of a network is the mean value of the cluster coefficient of its nodes. Therefore, the cluster coefficient measures the transitivity of the relationships in the graph. The value 1 (one) means that the relationships are all transitive, while the value 0 means that the relationships are all intransitive.

2.3 Academic social networks

By using social network analysis, researchers may understand and evaluate academic interactions in many ways. According to Melin and Persson (1996), using co-authorship we are able to study collaboration among researchers (the traditional co-authorship network), or to study institutional collaboration (when we analyze co-authorship among different institutions based on the author's professional address), or even international collaboration (co-author partnership among countries). Nevertheless, the authors also recognized that academic collaboration is something larger than co-authorship, once it can lead to other types of products and knowledge.

Mählck and Persson (2000) studied co-authorship and citation networks from two departments at different Swedish universities between 1986 and 1996. To analyze the co-authorship network the authors used the concept of socio-bibliometric maps, where nodes (authors) were labeled according to some status such as gender, academic degree, etc. and links (relationship) besides indicating co-authorship could also highlight if the persons had an advisor-student relationship. Among the results, the authors found that the most productive authors were PhDs, and they were surrounded by less productive ones, who were, mostly, their students.

Yousefi-Nooraie et al. (2008) analyzed the co-authorship networks of three Iranian Medical academic research centers to study its scientific productivity (articles written in English). As a result, authors found that centers with denser and more

decentralized networks, and that are also more open to outside connections had better scientific outcomes.

Bellotti (2012) studied how variations in network measures (in micro, i.e., collaboration between scientists; in macro level, that is, between institutions; and in meso level, that is, micro and macro metrics combined) could explain variations in the total money that an Italian Physicist receives to fund his/her research. As a result, the author inferred that researchers that collaborate with many different Physicists (i.e., that change partners over the years) tend to get more money. This characteristic was even more important than working in a big university or having many connections in the network.

Concerning to social network studies dealing (in some manner) with the P&S community, we shall highlight the work from Baccini, Barabesi and Marcheselli (2009). The authors studied editorial politics of Statistics & Probability journals creating a network where two journals are linked if they share a same editor in their boards. Moreover, the editorial proximity of two journals could be valued by the strength of the tie. The resulting network was very compact, which could be seen, according to the authors, as evidence of a common perspective about appropriate investigation methods and theoretical development in the domain of Probability and Statistics.

De Stefano, Giordano and Vitale (2011) critically discussed some issues in the analysis of co-authorship networks such as: data collection, network boundary setting, relational data matrix definition, data analysis and interpretation of results. Furthermore, authors illustrated their argumentation using real data based on researchers involved in four disciplines (Physics, Engineering, Arts & Humanities and Economics & Statistics) at the Italian university of Salerno.

Stefano et al. (2013) aimed to compare co-authorship network results of Italian academic statisticians using three data sources (Web of Science, Current Index to Statistics and nationally funded research projects). As a result, authors observed the small-world structure of the networks and for some statistic subfields they also found evidences that the authors seem to behave as if they are guided by a scale-free distribution. Furthermore, the general idea of positive association between statisticians' performance (h-index) and their central positions in the network was confirmed. However, some results may depend on the Bibliographic archives source.

As done by Abbasi, Altmann and Hossain (2011) and Cimenler, Reeves and Skvoretz (2014), Bordons et al. (2015) run a Poisson regression model to studied the relationship between the research performance (g-index) of scientists and his/her position in co-authorship network. Moreover, the authors analyzed three co-author ship networks (Nanoscience, Pharmacology and Statistics) in Spain during 2006 to 2008, to understand trends in each one of the fields. As a result, they found that Statistic Network was less dense, less connected and more fragmented than the others. The degree centrality and the strength of links were positive related

with the g-index in all three fields; however, the benefits (in terms of g-index) from the author position in the network were smaller in the Statistics field.

Said, Wegman and Sharabati (2010) proposed a model of preferential attachment in co-authorship networks and used it to predict emerging scientific subfields over time. They argued that the process of one actor attaching to another and strengthening the tie over time is a stochastic random process based on the distributions of tie-strength and clique size among authors. Thus, they used empirical data of statisticians working in prominent American Universities, focusing on the biopharmaceutical subfield, to estimate these distributions.

In Brazil, there are studies about co-authorship in several areas of knowledge. Mena-Chalco and Cesar (2009) developed a software named scriptLattes that extracts and analyzes data from the Curriculum Lattes, and it became an important and useful tool for those interested in academic network and bibliometric analysis. Using the scriptLattes, Mena-Chalco et al. (2014) were able to evaluate over one million curriculums of Brazilian researches. Andretta (2012) studied the scientific production of graduate programs in Information Science in Brazil, analyzing issues such as the profile of the production, productivity, and scientific collaboration, highlighting the characteristics of each Brazilian region. Andretta, Silva and Ramos (2012) repeated the same study focusing on the State of São Paulo. Alves, Yanasse and Soma (2014) evaluated the profile of CNPq's research productivity fellows in Chemistry in Brazil. Costa et al. (2013) investigated the scientific collaboration among the Brazilian northeast researchers working in biotechnology. These authors also identified which are the main universities in the region connected with foreign centers. Nascimento and Beuren (2011) studied the scientific production networks among graduate programs in Accounting in Brazil.

3 Brazilian probability and statistics dataset

Each research can register in his/her Lattes curriculum from zero to six expertise areas. For this, the areas in the Lattes Platform are represented by four levels of hierarchy, namely: major knowledge area; area; subarea; and specialty.

There are nine major knowledge areas that can be chosen by the researcher, as follows: Exact and Earth Sciences; Biological Sciences; Engineering; Health Sciences; Agricultural Sciences; Applied Social Sciences; Human Sciences; Linguistics, Letters and Arts; and Other. The Exact and Earth Sciences major area is divided into eight areas, namely: Mathematics; Probability and Statistics; Computer Science; Astronomy; Physics; Chemistry; Geosciences; and Oceanography.

The P&S area, in turn, is divided into three subareas: Statistics; Probability; and Applied Statistics and Probability. The Statistics subarea is divided into eight specialties: Data Analysis; Multivariate Analysis; Fundamentals of Statistics; Inference in Stochastic Processes; Non-Parametric Inference; Parametric Inference; Design of Experiments; Regression and Correlation. The Probability subarea is

divided into six specialties: Stochastic Analysis; Special Stochastic Processes; Markov Processes; Limit Theorems; General Theory and Probability Foundations; General Theory and Stochastic Processes. The Applied Probability and Statistics subarea is a specialty in itself.

Some authors, such as Arruda et al. (2009), argue that this division imposed by Lattes curricula structure is not clear for certain research purposes. In addition, certain knowledge areas (and their subsequent levels in the hierarchy) include issues related to P&S, for example, there is a specialty named Methods and Mathematical Models, Econometric and Statistics, in the Quantitative Methods in Economics subarea, in the Economics area that belongs to the major knowledge area of Applied Social Sciences. This sequence can easily be confused with the Probability and Applied Statistics subarea/specialty. Besides, authors can also include other subareas in the Lattes curricula if they do not find an adequate one to describe their research.

4 Method

The methodology used in this paper was organized in four activities: data gathering; sample selection; relevant information extraction; calculation of metrics.

In the *data gathering* activity, it was used the XML raw file from 3.2 millions of curricula from Lattes Platform.[5] These curricula were downloaded by the researchers from the Social Network Analysis and Scientometrics Group[6] in July 2013. This group aims to study the characteristics of the entire Brazilian scientific production and developed a methodology to obtain and organize the curricula files from Lattes Platform. More details see Digiampietri et al. (2014) and Mena-Chalco et al. (2014).

In the *sample selection* activity, from the 3.2 million curricula, were select for this study the ones that satisfy three criteria: curricula from PhDs (first criterion) that contain the "Probability and Statistics" value in the field "Areas of Expertise" (second criterion) and which professional address is in Brazil (third criterion). This activity identified 2,373 curricula.

In the *relevant information extraction* step, the following information from the Lattes curricula were extracted and organized: professional address (for geolocation of the curricula); expertise areas; relationships among curricula (the Lattes Curricula have explicit relationships information of coauthors, advisors, advisees, members of a scientific project, etc.).

In this data set of study, we observe that 2147 (91.61%) researchers were born in Brazil and the others are from 41 different nations; especially from Peru (48), Argentina (23), France (11) and Cuba, Colombia and the United States (10 PhDs

[5] http://lattes.cnpq.br/.

[6] http://dgp.cnpq.br/dgp/espelhogrupo/9125239221851493.

each). Moreover, 2226 PhDs (93.81%) have Brazilian nationality. Table 1 shows the distribution of PhDs with Brazilian nationality by Region. One can clearly see that the foreigners PhDs traditionally work in the Southeast region (69.39%).

Table 2 summarizes some characteristics of PhDs working with P&S by Region. There we can see that the Southeast is the region that concentrates the majority (over 58%) of PhDs working in P&S in Brazil; on the other hand, the North region concentrates only 3% of them. Regarding to advisorship activities, the PhDs form the Midwest had advised on average 6.64 graduation students; PhDs form Northeast had advised on average 4.21 Scientific initiation students; PhDs from South had advised on average 3.24 specialization students and 4.99 master dissertations; while Southeast PhDs had advised on average 4.62 doctoral theses.

In addition to these explicit relationships, we also used the algorithm presented in Digiampietri et al. (2012) for the identification of coauthorship relationships that were not explicitly present in the curricula (the absence of this information on the curricula occurs, typically, due to the lack of standardization in the filling of the name of the authors in the publications' registers). All these relationships (explicit

Table 1 *Distribution of PhDs working with P&S by nationality and Brazilian region*

Region	Brazilian	Foreign	Total
North	68	4	72
Northeast	338	12	350
Central-West	179	13	192
Southeast	1276	102	1378
South	365	16	381
Brazil	2226	147	2373

Table 2 *Some informations of PhDs researchers working with P&S by Brazilian region*

Region	PhDs %	Graduation students advised	Scientific initiation students advised	Specialization students advised	Master students advised	PhD students advised	Others tips of advisorship
North	3.03	4.21	3.51	1.67	2.82	0.22	0.54
Northeast	14.75	3.60	4.21	1.58	3.84	0.60	1.79
Central-West	8.09	6.64	3.73	1.66	3.04	0.38	1.79
Southeast	58.07	3.89	2.84	1.08	4.62	4.62	2.18
South	16.06	5.07	3.59	3.24	4.99	1.09	2.46
Brazil	100	4.27	3.26	1.56	4.38	1.13	2.08

or not) were used in the production of the academic social networks. After that, 29 social networks (graphs) were produced: one composed of the 2373 researchers from the sample, and 27 additional networks composed only of researchers from each one of the Brazilian States and the Federal District and one network considering each state as a node.

In the *calculation of metrics* activity, we measured the social networks structural metrics (see Wasserman and Faust (1994)). These metrics aim to explain some characteristics from the networks to allow their understanding and comparison.

5 Results

Based on the relationships from the 2373 curricula, 29 academic social networks were produced. These networks will be presented and analyzed in this section.

5.1 Data description

We first describe characteristics of the PhDs in our sample according to their location (Table 3) and expertise (Table 4).

The Brazilian government estimates its population at 202 million people (The Brazilian Institute of Geography and Statistics—IBGE[7]). More than half (55%) of the Brazilian population is located in the five most populous states (respectively, SP, MG, RJ, BA and RS). These five states contains 66% of the PhDs working with P&S. The variables *Percentage of PhDs Working with P&S* and *Average time since PhD graduation* are highly correlated (the value of the Pearson correlation is 0.544 with p-value < 0.05). The most populous state (SP) has the highest *Average time since PhD graduation* (13.9 years) and the three less populous have the lowest values for these variable (AP, AC and RR, with 4, 6.3 and 6.5 years, respectively). The number of *PhD working with P&S in Brazil per million people* ranges from 1 (in AP) to 31 (in DF). On average, there are 12 PhDs working with P&S in Brazil per million people. The highest concentration of PhDs per million people occurs in the Federal District (DF) which contains universities and federal agencies that employ many of these PhDs.

Regarding the expertise of the PhDs working with P&S, in Table 4, we can see that two subareas concentrate the great majority of such PhDs: Applied Probability and Statistics (41.84%) and Statistics (38.31%). Therefore, approximately 8 out of 10 PhDs working with P&S areas, work in at least one of these subareas. In third place, there are 7.82% of such PhDs working in the Probability subarea. This suggests that the great majority of the PhDs working with P&S develop more applied than theoretical researches.

[7] http://www.ibge.gov.br/home/estatistica/populacao/estimativa2014/.

Table 3 *Some informations of the PhDs working with P&S by Brazilian states*

State	Population	Pop. %	PhDs working with P&S	PhDs working with P&S %	Average time since PhD graduation	PhDs working graduation with P&S per million people
AC	790,101	0.39%	4	0.17%	6.3	5
AL	3,321,730	1.64%	17	0.72%	10.8	5
AM	3,873,743	1.91%	20	0.84%	11.4	5
AP	750,912	0.37%	1	0.04%	4.0	1
BA	15,126,371	7.46%	75	3.16%	9.3	5
CE	8,842,791	4.36%	52	2.19%	9.8	6
DF	2,852,372	1.41%	87	3.67%	12.3	31
ES	3,885,049	1.92%	37	1.56%	8.1	10
GO	6,523,222	3.22%	44	1.85%	8.3	7
MA	6,850,884	3.38%	12	0.51%	8.6	2
MG	20,734,097	10.22%	281	11.84%	10.1	14
MS	2,619,657	1.29%	30	1.26%	10.2	11
MT	3,224,357	1.59%	31	1.31%	8.1	10
PA	8,073,924	3.98%	32	1.35%	8.8	4
PB	3,943,885	1.94%	50	2.11%	8.5	13
PE	9,319,347	4.60%	76	3.20%	11.8	8
PI	3,194,718	1.58%	5	0.21%	12.2	2
PR	11,081,692	5.46%	145	6.11%	9.9	13
RJ	16,461,173	8.12%	378	15.93%	11.8	23
RN	3,408,510	1.68%	50	2.11%	9.4	15
RO	1,748,531	0.86%	8	0.34%	7.8	5
RR	496,936	0.25%	2	0.08%	6.5	4
RS	11,207,274	5.53%	160	6.74%	11.0	14
SC	6,727,148	3.32%	76	3.20%	11.9	11
SE	2,219,574	1.09%	13	0.55%	10.3	6
SP	44,035,304	21.71%	682	28.74%	13.9	15
TO	1,496,880	0.74%	5	0.21%	8.2	3
Brazil	202,810,182	100.00%	2373	100.00%	11.5	12

5.2 The academic social network of researchers in probability and statistics

We will start our analysis focusing in the Brazilian state network, in which each state (including the Federal District-DF) is treated as a node, as shown in Figure 1. The nodes were disposed in the graph in order to arrange those with higher centrality measures in the center and the ones with lower measures in the periphery. In this context, four central states should be highlighted: SP, MG, RJ and PE. According to the CAPES's triennial[8] assessment report 2013, those are the unique

[8] http://www.avaliacaotrienal2013.capes.gov.br/relatorios-de-avaliacao.

Table 4 *Distribution of PhDs according with their expertise*

Subarea	Percentage
Applied Probability and Statistics	41.84%
Statistics	38.31%
Probability	7.82%
Biostatistics	0.95%
Time Series	0.44%
Applied Statistics	0.36%
Spatial Statistics	0.36%
Multivariate Statistics	0.28%
Bayesian Inference	0.28%
Econometrics	0.24%
Experimental Statistics	0.24%
Stochastic Processes	0.24%
Other	8.65%

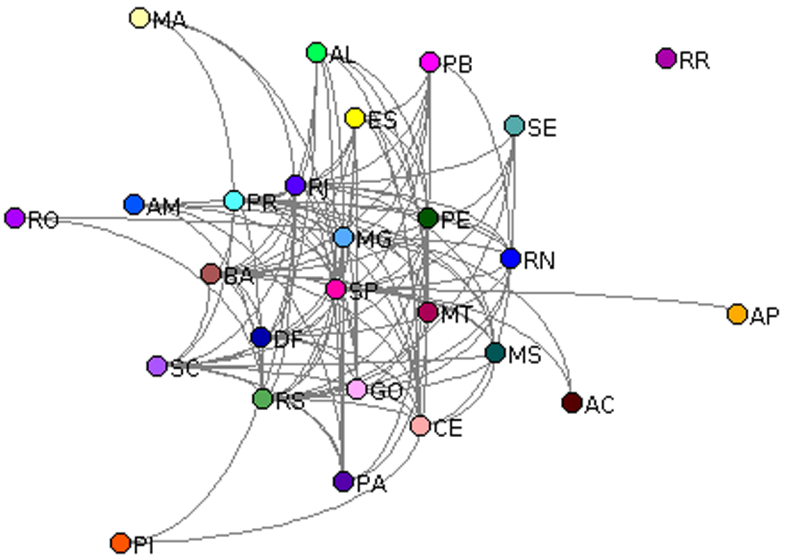

Figure 1 *P&S collaboration network—Brazilian states network.*

Brazilian states with Statistic doctoral programs (SP–USP, UNICAMP and UFSCAR; MG–UFMG; RJ–UFRJ; PE–UFPE).

We can see in Figure 1 that the network is disconnected because of RR. Furthermore, its diameter is equal to 3 and its average path length is 1.63. These two metrics indicate the traditional small word idea (see Travers and Milgram (1969) and Milgram (1967)).

Table 5 *Global metric of P&S collaboration network—Brazilian states network*

Metric	Value
Clique Number	8.00
Average Path Length	1.63
Clustering Coefficient	0.59
Centralization degree	0.52
Centralization closeness	0.23
Centrelization evcent	0.52
Diameter	3.00
Graph density	0.41

The average degree is 11.48 and the average cluster coefficient 0.59. Jackson (2008) pointed that a feature of social networks is that they tend to have high cluster coefficient when compared to random networks. For example, a random network with 27 nodes and an average degree of 11.48 has a cluster coefficient of about 0.43 (11.48/27), which corroborates this trend. Other metrics are summarized in Table 5.

Table 6 exposes four centrality metrics (degree, betweenness, closeness and eigenvector) of each state. Clearly, the most central states are from the Southeast region, specially, SP, MG and RJ. Moreover, when one observes the centralization metrics in Table 5, particularly the centralization degree and centralization eigenvector, it is possible to understand the importance of SP as the central and most important and connected vertex of the network. On the other hand, the states from the north region were the less central.

In the network from Figure 2, each node represents a Brazilian city (in which there is at least one PhD working with P&S). The edges between cities indicate there is at least one academic collaboration between the PhDs from these cities. Edges between cities from the same state are colored, and the ones between cities from different states are gray. It is possible to observe a concentration of nodes (cities) in the states from the South and Southeast regions. On the other hand, in the North and Central-West regions there are few cities with PhDs working with P&S.

In the network from Figure 3, each node represents one PhD working with P&S. Each node was positioned in the Brazilian Map close to its professional address (in order to minimize overlays, the nodes were not positioned exactly over their professional address). The network's edges correspond to the relationship between two PhDs (nodes). In this figure, it is possible to observe the concentration of nodes in the states' capitals and, as seen in Figure 2, a concentration of nodes in the states from the South and Southeast regions, followed by the states in the Northeast region. It is worth to note that there is a greater concentration of nodes in and around the cities which host P&S graduate programs, namely: Brasília, Belo

Table 6 *Centrality metrics of P&S collaboration network by Brazilian states*

State	Betweenness	Closeness	Degree	Eigenvalue
AC	0.000	0.013	4	0.155
AL	0.002	0.014	10	0.499
AM	0.000	0.014	7	0.325
AP	0.000	0.013	3	0.079
BA	0.013	0.016	16	0.781
CE	0.042	0.016	16	0.728
DF	0.006	0.015	13	0.644
ES	0.007	0.015	13	0.648
GO	0.004	0.014	10	0.461
MA	0.000	0.013	5	0.207
MG	0.121	0.018	23	0.966
MS	0.003	0.015	11	0.544
MT	0.003	0.014	11	0.538
PA	0.003	0.014	10	0.469
PB	0.001	0.014	9	0.448
PE	0.027	0.016	18	0.835
PI	0.000	0.013	4	0.123
PR	0.032	0.016	17	0.785
RJ	0.092	0.017	20	0.840
RN	0.009	0.015	13	0.621
RO	0.000	0.013	4	0.143
RR	0.000	0.001	2	0.000
RS	0.058	0.017	19	0.850
SC	0.004	0.015	12	0.599
SE	0.000	0.014	7	0.326
SP	0.199	0.019	25	1.000
TO	0.001	0.014	8	0.353

Horizonte, Belém, Recife, Rio de Janeiro, Natal, São Paulo, Campinas and São Carlos.

Figure 4 aims to clarify the relationships among the PhDs working with P&S. The network presented in this figure corresponds to two modifications applied in the network from Figure 3: nodes with degree zero were excluded; and the nodes were reorganized to approximate the nodes that are related and to move away the nodes that are not related. It is possible to observe in the center of Figure 4 the giant component. It is composed by the majority of the nodes of this network. Surrounding this component, there are dozens of smaller components, most composed only by two or three nodes.

Table 7 presents the distribution of the number of connected components from Figure 4 according to their size. As observed, the network contains a giant component with 1197 nodes and several small components (composed by two to seven nodes). It is important to observe that almost more than one thousand PhDs (about 44% of PhDs studied in this paper) do not have any relationship with other PhD

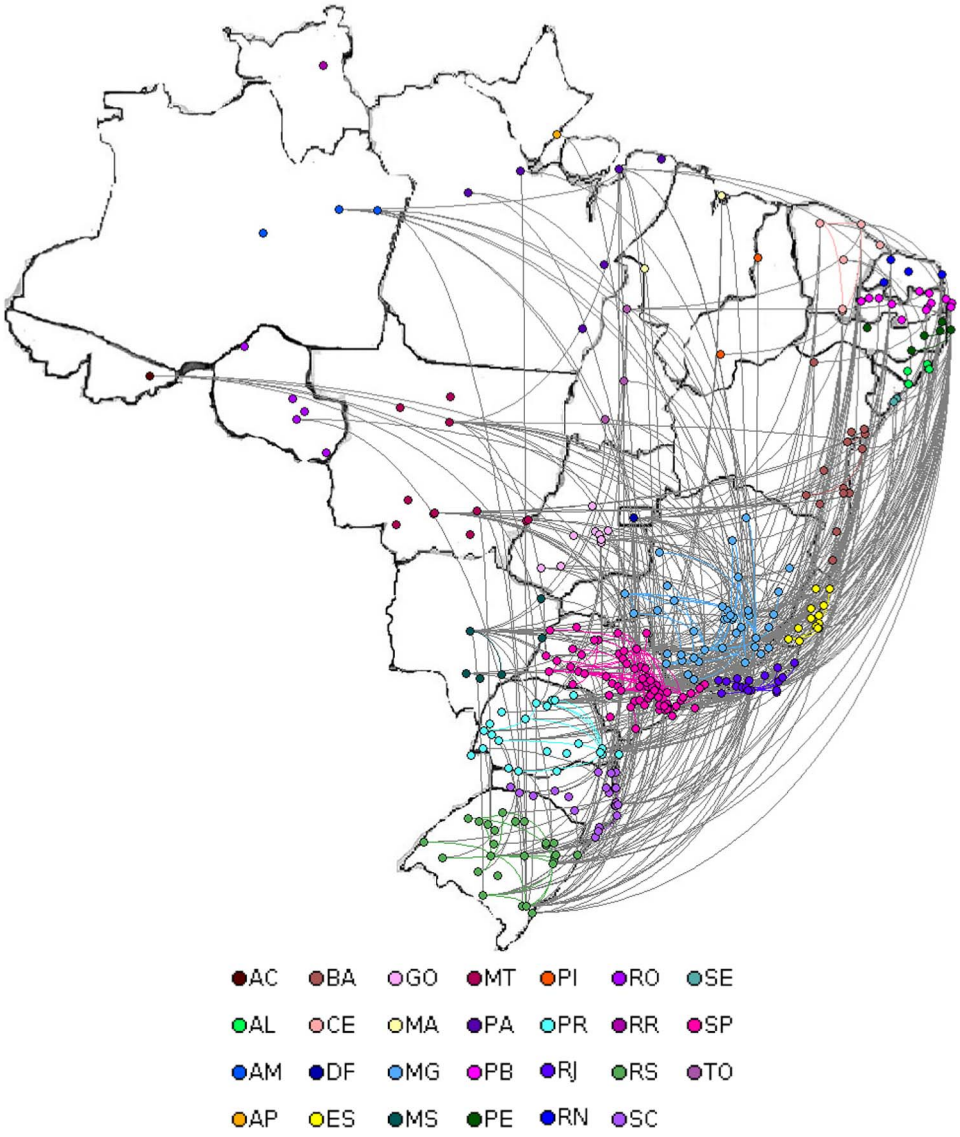

Figure 2 *Probability and Statistics Collaboration Network—Brazilian cities.*

from the sample. These PhDs were not plotted in the network presented in Figure 4.

Figure 5 presents the percentage of edges between the PhDs from each state. Each line in the table sums 100%, corresponding to all the relationships from the respective state. The penultimate column presents the total number of links from the respective state. Furthermore, the last column presents the E-I index.

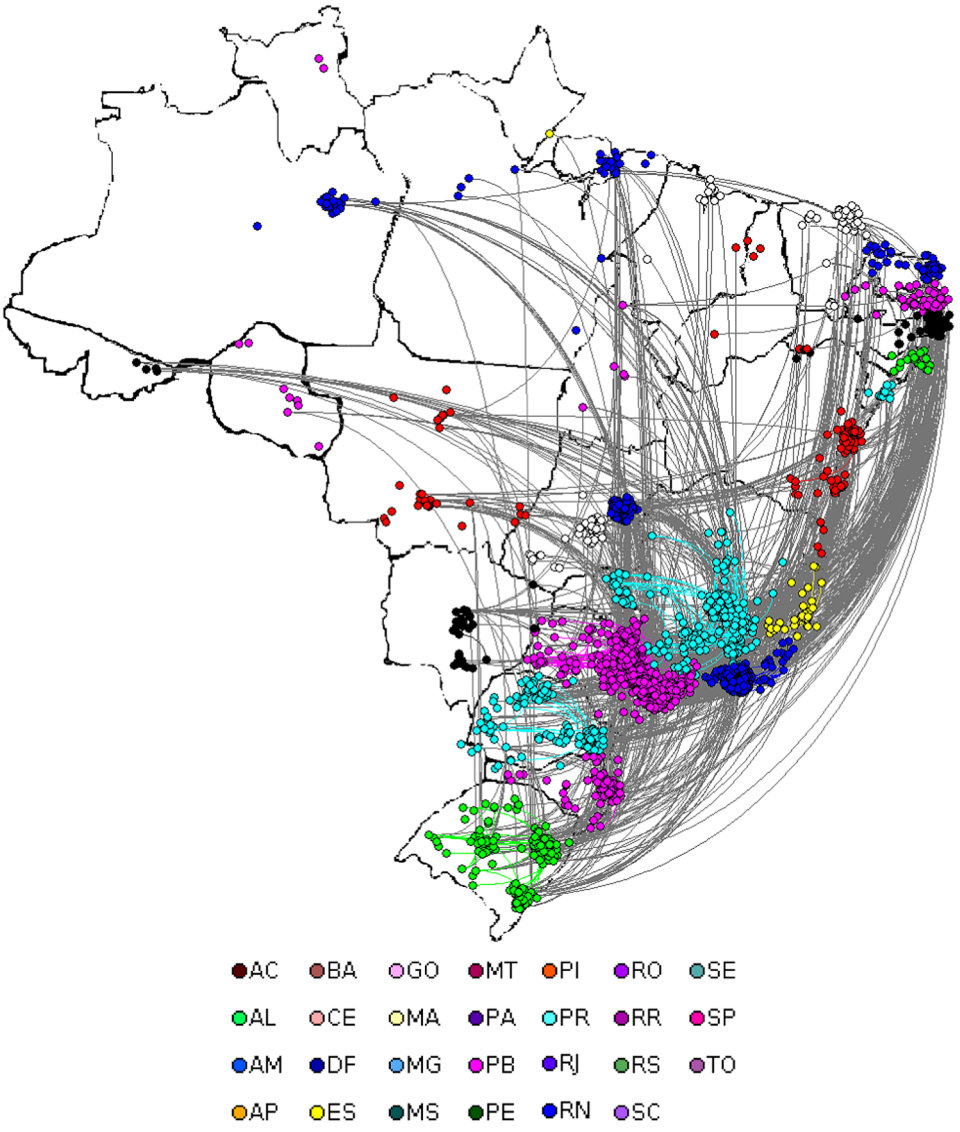

Figure 3 *Probability and Statistics Researchers Collaboration Network distributed according to their professional address.*

Figure 6 presents the degree distribution of the Brazilian network. Over 40% of the nodes have degree zero, and less than 10% of the researchers have degree equal to ten or higher, with only one PhD from PE with degree equal to 87, indicating the rich-get-richer idea (Easley and Kleinberg (2010)). Additionally, still based on the degree distribution, we use the Kruskal–Wallis test to investigate regional degree differences. Figure 7 shows the distribution of the ln(degree + 1)

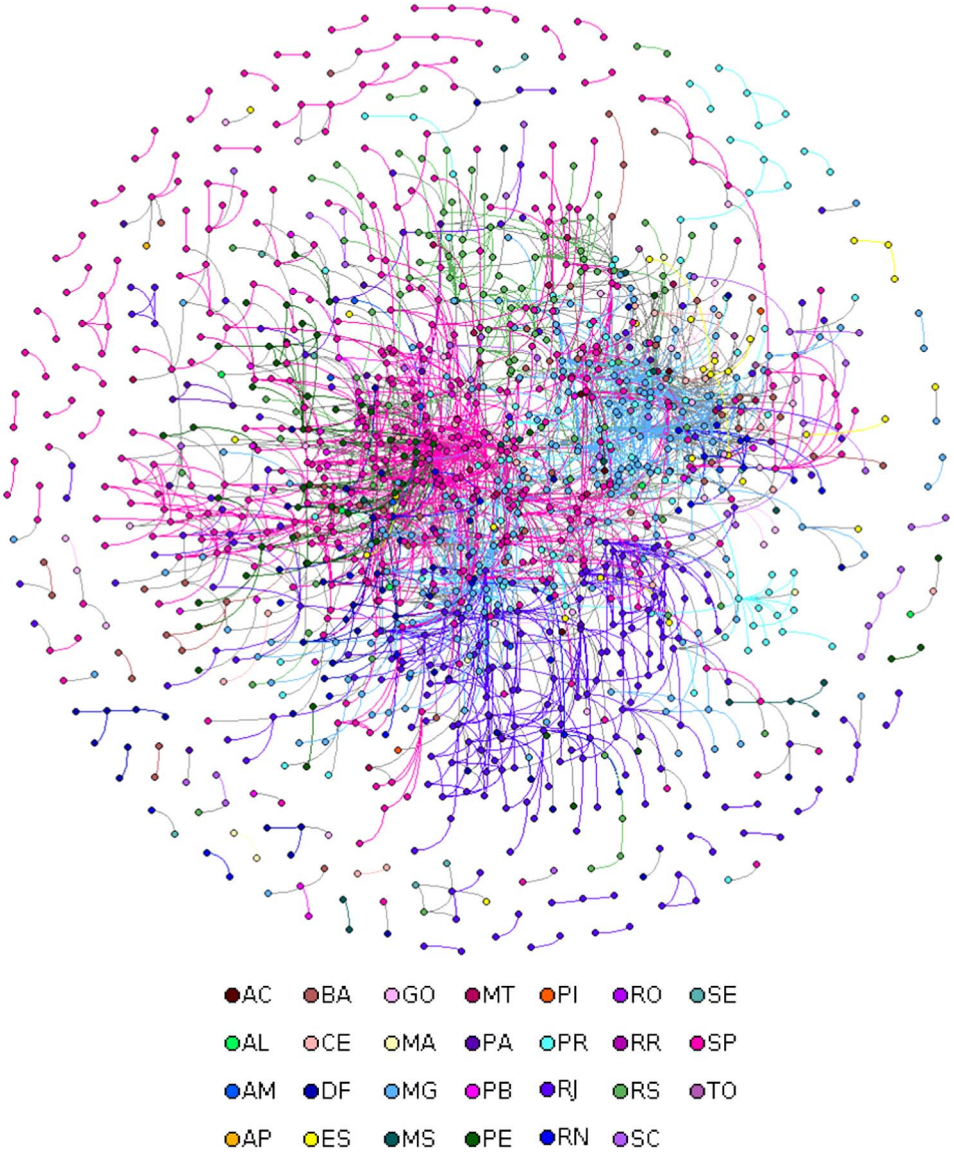

Figure 4 *Probability and Statistics Researchers Collaboration Network—reorganized.*

for each Brazilian region. With a p-value of 0.0001 (KW-H(4, 2373) = 22.6793) there are statistical evidences against the null hypothesis, i.e., we can detect differences in degree distribution by region, specially between North and Southeast regions.

Table 7 *Number and size of the connected components*

Component size	Number of components
1	1036
2	54
3	10
4	7
5	3
6	1
7	1
1197	1

Assuming each Brazilian state as a group[9] in the network, we are able to calculate the E-I index for each one of them.[10] Most of the states had positive E-I index, with the exception of Rio de Janeiro (RJ) and São Paulo (SP). This result seems natural since these states are the most developed ones in Brazil, and researchers in these locations can find partners more easily in their own state. On the other hand, states from the north and northeast seem to have a greater degree of external dependence because the E-I index in these states are close to 1, or even equal to 1.

Still according to Figure 5, the cell's background is colored according to its value (higher values implies in a higher green tonality). It is worth to highlight three different information from this table. The first indicates the importance of some states concerning to relationships (the states whose columns have more green cells), especially São Paulo (SP) and Minas Gerais (MG), followed by Pernambuco (PE) and Rio de Janeiro (RJ). The second is the percentage of self-relationships (relationships between nodes from the same state), in this criteria, the most important states are: São Paulo (SP), Rio de Janeiro (RJ), Rio Grande do Sul (RS) and Minas Gerais (MG). At last, it is possible to observe that most of the remaining green cells corresponds to relationships between states geographically close, for example, 50% of the relationships involving PhDs (nodes) from Piauí (PI) occur with PhDs from Ceará (CE); 20% of the relationships involving nodes from Maranhão (MA) occur with nodes from Pará (PA); and 26% from the relationships involving nodes from Alagoas (AL) and Paraíba (PB) occurs with PhDs from Pernambuco (PE) (neighboring states).

Table 8 contains the network metrics that were measured in both the national and the states level. The networks are sorted according to the number of nodes, and the ten biggest state networks are highlighted. Following, some characteristics of these networks will be discussed, focusing on the biggest networks. It is possible to observe a disparity in the size of the networks: the five smallest networks contains,

[9] We may think in a particular state as the *IG*, and the others as the *EG*, for example.

[10] It is worth noting that, as the researchers from Roraima (RR) does not have internal or external links in the network, it was not possible to calculate the E-I index for RR.

	AC	AL	AM	AP	BA	CE	DF	ES	GO	MA	MG	MS	MT	PA	PB	PE	PI	PR	RJ	RN	RO	RR	RS	SC	SE	SP	TO	Total	E-I index
AC	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	62%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	31%	0%	13	0.85
AL	0%	4%	0%	0%	4%	4%	4%	0%	0%	0%	19%	0%	4%	0%	0%	26%	0%	0%	11%	0%	0%	0%	0%	0%	0%	26%	0%	27	0.93
AM	0%	0%	9%	0%	0%	0%	0%	0%	0%	0%	43%	0%	0%	4%	0%	0%	0%	0%	9%	0%	0%	0%	4%	0%	0%	30%	0%	23	0.83
AP	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	1	1.00
BA	0%	1%	0%	0%	16%	2%	1%	1%	0%	0%	27%	1%	3%	0%	1%	10%	0%	4%	6%	0%	0%	0%	2%	1%	0%	24%	0%	100	0.68
CE	0%	2%	0%	0%	3%	20%	3%	2%	0%	0%	3%	3%	0%	0%	0%	6%	2%	3%	0%	12%	0%	0%	6%	8%	0%	26%	3%	66	0.61
DF	0%	1%	0%	0%	1%	2%	29%	0%	2%	0%	19%	0%	0%	0%	0%	2%	0%	0%	11%	7%	0%	0%	2%	1%	0%	25%	0%	121	0.42
ES	0%	0%	0%	0%	2%	2%	0%	16%	2%	0%	42%	0%	2%	0%	2%	6%	0%	3%	13%	0%	0%	0%	3%	0%	0%	8%	0%	62	0.68
GO	0%	0%	0%	0%	0%	0%	5%	2%	11%	0%	16%	0%	0%	0%	0%	7%	0%	7%	0%	2%	0%	0%	0%	0%	0%	48%	2%	44	0.77
MA	0%	0%	0%	0%	0%	0%	0%	0%	0%	20%	0%	0%	0%	0%	0%	0%	0%	20%	20%	0%	0%	0%	0%	0%	0%	40%	0%	5	0.60
MG	1%	1%	1%	0%	4%	0%	3%	3%	1%	0%	44%	1%	2%	0%	1%	2%	0%	4%	7%	1%	0%	0%	4%	1%	0%	16%	1%	748	0.12
MS	0%	0%	0%	0%	3%	5%	0%	0%	0%	0%	21%	11%	0%	0%	0%	0%	0%	11%	3%	0%	0%	0%	5%	3%	0%	34%	5%	38	0.79
MT	0%	2%	0%	0%	6%	0%	0%	2%	0%	0%	33%	0%	8%	2%	0%	2%	0%	6%	0%	0%	0%	0%	4%	0%	0%	35%	0%	49	0.84
PA	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	9%	0%	3%	24%	0%	9%	0%	0%	6%	0%	0%	0%	3%	12%	0%	32%	0%	34	0.53
PB	0%	0%	0%	0%	2%	0%	0%	2%	0%	0%	11%	0%	0%	0%	15%	40%	0%	0%	2%	5%	0%	0%	0%	0%	0%	24%	0%	55	0.71
PE	0%	3%	0%	0%	5%	2%	1%	2%	1%	0%	6%	0%	0%	1%	10%	29%	0%	2%	2%	1%	0%	0%	2%	0%	0%	31%	0%	217	0.43
PI	0%	0%	0%	0%	0%	50%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	50%	0%	0%	0%	0%	0%	0%	0%	0%	2	1.00
PR	0%	0%	0%	0%	2%	1%	0%	1%	1%	0%	15%	2%	1%	0%	0%	2%	0%	33%	5%	1%	0%	0%	9%	4%	0%	22%	1%	221	0.34
RJ	0%	1%	0%	0%	1%	0%	3%	2%	0%	0%	12%	0%	0%	0%	0%	1%	0%	2%	55%	1%	0%	0%	2%	0%	1%	17%	0%	447	-0.11
RN	0%	0%	0%	0%	0%	11%	11%	0%	1%	0%	15%	0%	0%	0%	4%	3%	0%	3%	5%	28%	0%	0%	3%	0%	1%	15%	0%	74	0.43
RO	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	75%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	25%	0%	0%	0%	0%	4	1.00
RR	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0	-
RS	0%	0%	0%	0%	1%	2%	1%	1%	0%	0%	12%	1%	1%	0%	0%	2%	0%	9%	3%	1%	0%	0%	48%	7%	0%	11%	0%	234	0.04
SC	0%	0%	0%	0%	1%	6%	1%	0%	0%	0%	7%	1%	0%	4%	0%	0%	0%	10%	2%	0%	0%	0%	18%	29%	0%	21%	0%	90	0.42
SE	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	11%	0%	0%	33%	11%	0%	0%	11%	0%	11%	22%	0%	9	0.78
SP	0%	1%	1%	0%	2%	1%	2%	0%	2%	0%	10%	1%	1%	1%	1%	6%	0%	4%	6%	1%	0%	0%	2%	2%	0%	55%	0%	1217	-0.10
TO	0%	0%	0%	0%	0%	12%	0%	0%	6%	0%	47%	12%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	0%	12%	0%	17	1.00

Figure 5 *Percentage of relationships among states.*

together, less than 0.75% of the total number of nodes. On the other hand, the São Paulo network contains more than 28% of the nodes, and, adding the nodes from Rio de Janeiro and Minas Gerais they represent more than half of the total number of PhDs working in Brazil with P&S. Together, the ten biggest state networks contain 85% of the nodes.

The third column from Table 8 contains the number of nodes with degree greater than zero (i.e., the nodes that have at least one relationship) in each network. In the national network, only 1391 from the 2373 PhDs have at least one relationship. The calculation of the remainder metrics presented in this table used only the nodes with degree greater than zero. The fourth column presents the number of edges in each network. The national network contains 2791 edges. It is worth to observe that, despite being the third greatest network, the Minas Gerais network is the second one in the number of edges. This characteristic influences several metrics as will be presented as follows.

As explained in Section 2.2, the size of the giant component corresponds to the number of nodes in the biggest connected component in each network. In this work, the percentage of nodes in the giant component was calculated considering only the nodes with degree greater than zero. Giant components with a high percentage of the network's nodes are considered a positive aspect in a social network. This indicates that a significant number of individuals belongs to the main flow of information/knowledge in the network. In the national network, 86.1% of the nodes with degree greater than zero are in the giant component. Among the biggest state networks, stands out the Minas Gerais network with 87.8% of its nodes in its giant component. Based on Density measure, all networks assessed

Table 8 *Networks' metrics*

State	Nodes	Nodes with degree greater than zero	Edges	Nodes in the giant component	% of nodes in the giant component	Density	Average degree	Clustering coefficient	Degree centralization	Closeness centralization	Betweeness centralization	Diameter	Maximum clique size	Average path length	Eigenvector centralization
AP	1	0	0	1	–	–	–	–	–	–	–	0	1	–	–
RR	2	0	0	1	–	–	–	–	–	–	–	0	1	–	–
AC	4	2	1	2	100.00	1.000	1.000	–	0.167	0.139	0.000	1	2	1.000	1.000
PI	5	0	0	1	–	–	–	–	–	–	0.000	0	1	–	–
TO	5	0	0	1	–	–	–	–	–	–	0.000	0	1	–	–
RO	8	0	0	1	–	–	–	–	–	–	0.000	0	1	–	–
MA	12	2	1	2	100.00	1.000	1.000	–	0.076	0.014	0.000	1	2	1.000	1.000
SE	13	2	1	2	100.00	1.000	1.000	–	0.071	0.012	0.000	1	2	1.000	1.000
AL	17	2	1	2	100.00	1.000	1.000	–	0.055	0.007	0.000	1	2	1.000	1.000
AM	20	3	2	3	100.00	0.667	1.333	0.000	0.095	0.010	0.006	2	2	1.333	0.977
MS	30	6	4	4	66.70	0.267	1.333	0.000	0.094	0.007	0.007	2	2	1.429	0.974
MT	31	6	4	4	66.70	0.267	1.333	0.000	0.058	0.006	0.004	3	2	1.571	0.957
PA	32	10	8	4	40.00	0.178	1.600	0.500	0.048	0.005	0.004	3	3	1.417	0.967
ES	37	16	10	5	31.30	0.083	1.250	0.000	0.096	0.005	0.009	2	2	1.410	0.971
GO	44	9	5	3	33.30	0.139	1.111	0.000	0.041	0.002	0.001	2	2	1.167	0.990
PB	50	9	8	7	77.80	0.222	1.778	0.231	0.096	0.005	0.011	3	3	1.818	0.964
RN	50	19	21	8	42.10	0.123	2.211	0.500	0.064	0.005	0.005	5	3	1.915	0.924

Table 8 *Continued*

rotatebox-90State	Nodes	Nodes with degree greater than zero	Edges	Nodes in the giant component	% of nodes in the giant component	Density	Average degree	Clustering coefficient	Degree centralization	Closeness centralization	Betweeness centralization	Diameter	Maximum clique size	Average path length	Eigenvector centralization
CE	52	13	13	5	38.50	**0.167**	2.000	**0.833**	0.049	0.003	0.003	3	4	**1.390**	0.960
BA	75	23	16	6	26.10	0.063	1.391	0.300	**0.035**	0.002	**0.002**	4	3	1.679	0.976
SC	76	29	26	9	31.00	0.064	1.793	0.255	0.084	0.002	0.008	3	3	1.783	0.967
PE	76	42	62	33	78.60	0.072	2.952	0.295	0.165	0.011	0.094	6	4	2.809	0.910
DF	87	31	35	19	61.30	0.075	2.258	0.270	0.107	0.005	0.026	5	4	2.370	0.939
PR	145	60	73	41	68.30	0.041	2.433	0.309	0.083	0.004	0.041	8	5	3.818	0.960
RS	160	76	112	63	82.90	0.039	2.947	0.307	0.054	0.005	0.043	11	4	4.398	0.893
MG	281	172	330	151	87.80	0.022	3.837	0.261	0.084	0.004	0.077	11	6	4.399	0.952
RJ	378	183	248	135	73.80	0.015	2.710	0.218	0.039	0.002	0.046	12	6	4.893	0.977
SP	682	384	671	291	75.80	0.009	3.495	0.239	0.037	0.001	0.042	18	5	5.714	0.972
Brazil	2373	1391	2791	1197	86.10	0.003	4.013	0.155	0.036	0.001	0.032	17	6	5.547	0.963

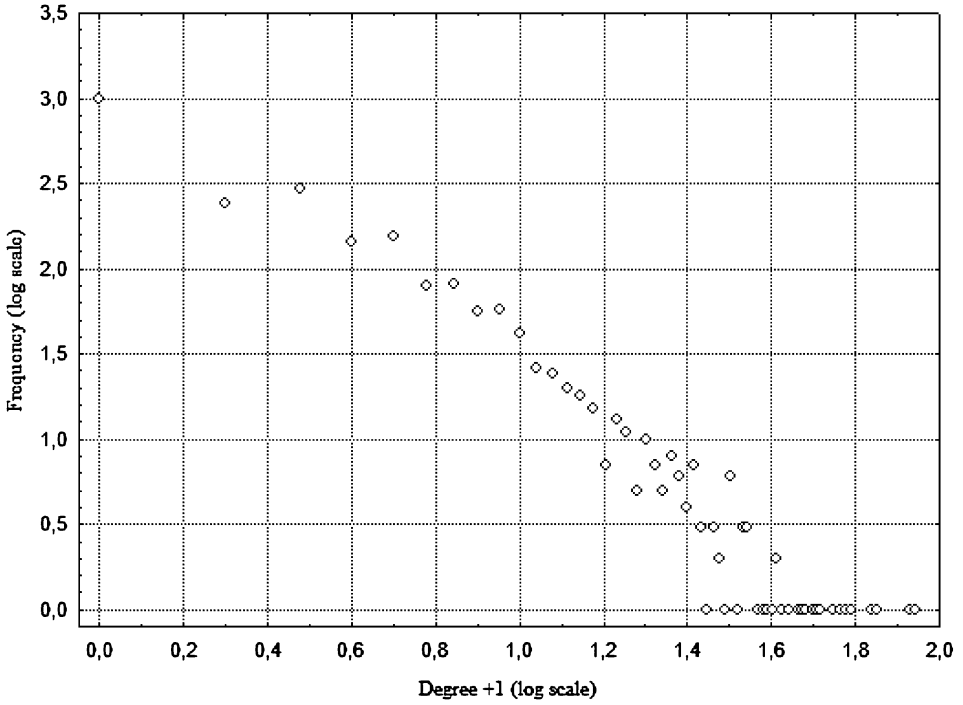

Figure 6 *Log-log degree distribution—Brazilian Network.*

in this paper are far to be complete.The density of the national network is 0.003 (i.e., there are only 1/333 of the number of possible edges for this network). The density of Ceará network is 0.167 (one sixth of the maximum number of edges for this network).

In Table 8 for all states, the average degree is greater than or equal to one. In the national network, the average degree is 4.01. In the biggest state networks, this value ranges from 1.39 (Bahia network) to 3.84 (Minas Gerais network). Still according to Table 8 the Ceará network (CE) stands out for having a high clustering coefficient (0.833). This value indicates that the Ceará network is cohesive. The clustering coefficient of the national network is 0.155 which suggests that this network is not cohesive and, probably, it is in a maturing stage.

The centrality metrics indicate how influential are the nodes in a network. High centrality values used to be viewed with caution in social networks, because they indicate that one individual is very important for the network. Thus, its eventual absence could imply in a great prejudice for the network. In the national network, the degree and closeness centralities have low value (0.036 and 0.001, respectively). In the state networks, the highest centrality value occurs in the Pernambuco network, and the lowest in the Bahia network (degree centrality) and São Paulo (closeness centrality).

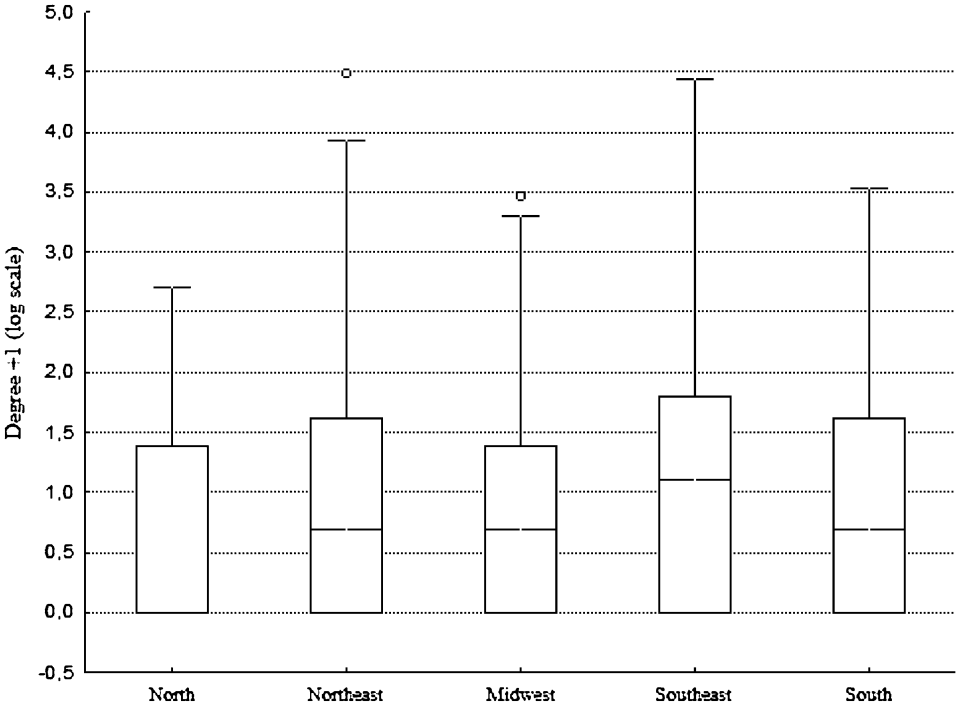

Figure 7 *Distribution of* log(degree + 1) *by Brazilian region.*

The diameter from Ceará and Santa Catarina networks is only three. The São Paulo network's diameter is 18. The diameter from the national network is 17, this value lower than the one from São Paulo network means in the national network there are shorter paths linking two persons from the same state (e.g., there are two PhDs in São Paulo which are not linked and do not have any common neighbor in the São Paulo network but they have a common neighbor in Mato Grosso do Sul network). The diameter in a social network used to be associated with the maximum time required to an information to be propagated to all the individuals from the network. Thus, lower values from this metric allows a faster information (or knowledge) propagation.

A clique with a great amount of people typically represents a cohesive group of PhDs (for example, a research group) that works in some particular area/subarea of expertise. The size of the maximum clique in the national, Rio de Janeiro, and Minas Gerais networks is six. On the other hand, in Ceará and Santa Catarina networks this value is only three.

The average path length corresponds to the average distance of the shortest path between all pairs of individuals in a connected component. This metric is also related to information (or knowledge) propagation speed in the network. In the national network, in average, the path between two PhDs is composed of 5.55

persons. In the Ceará network, this value is only 1.39. In the São Paulo network, the average path length is 5.71.

6 Final remarks

In this paper, we performed a social network analysis of PhDs working in the field of Probability and Statistics in Brazil, where ties represent either co-authorship, participation in joint project or the advisor-advisee relationship. Particularly, 29 networks were analyzed one for each state, one considering each state as a node and one for the whole country. As a result, the first network had small world characteristic, and the most central nodes were the states that host P& S doctoral programs. Regional differences were also detected. The biggest networks were form southeast and the smaller were from the north region. The same characteristic was observed with respect to the degree distribution. The national network shows that there is a greater concentration of nodes in and around cities having graduate programs in Probability and Statistics, which is also reflected in the size of the state networks. The clustering coefficient of the national network suggests that this community is not cohesive and, probably, it is in a maturing stage. Moreover, the E-I index indicated that states from the north and northeast have a greater dependence on collaboration with researchers from other states.

For further studies, we intend to investigate how network metrics can impact in productivity measures, which will allow us to use more sophisticated statistical methods in the (social) network analysis, as done by Abbasi, Altmann and Hossain (2011), Cimenler, Reeves and Skvoretz (2014), Bellotti (2012), de Arruda et al. (2013) and Peron, Costa and Rodrigues (2012).

References

Abbasi, A., Altmann, J. and Hossain, L. (2011). Identifying the effecs of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance mesuares and social network analysis mesuares. *Journal of Informetrics* **5**, 594–607.

Alves, A. D., Yanasse, H. H. and Soma, N. Y. (2014). Perfil dos bolsistas PQ da Área de Química baseado na Plataforma lattes. *Química Nova* **37**, 377–383.

Andretta, P. I. (2012). Uma análise sobre a produção, produtividade e colaboração na ciência da informação no Brasil entre os anos 2007 a 2009. *Palabra Clave* **1**, 48–52.

Andretta, P. I., Silva, E. and Ramos, R. (2012). Aproximações sobre produção, produtividade e colaboração científica entre os departamentos de ciência da informação do estado de São Paulo. *RDBCI* **9**, 46–63.

Ara, A. and Louzada, F. (2012). Descrição de algumas das dimensões que compõem o perfil do corpo docente dos departamentos de estatística do Brasil. *Boletim de Educação Matemática* **26**(42A), 23–38.

Arruda, D., Bezerra, F., Neris, V., Rocha De Toro, P. and Wainera, J. (2009). Brazilian computer science research: Gender and regional distributions. *Scientometrics* **79**, 651–665.

Baccini, A., Barabesi, L. and Marcheselli, M. (2009). How are statistical journals linked? A network analysis. *Chance* **22**, 35–45.

Bellotti, E. (2012). Getting funded. Multi-level network of physicists in Italy. *Social Networks* **34**, 215–229.

Bojanowski, M. and Corten, R. (2014). Measuring segregation in social networks. *Social Networks* **39**, 14–32.

Bonacich, P. and Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks* **23**, 191–201.

Bordons, M., Aparicio, J., González-Albo, B. and Díaz-Faes, A. A. (2015). The relationship between the research performance of scientists and their position in co-authorship networks in three fields. *Journal of Informetrics* **9**, 135–144.

Cimenler, O., Reeves, K. A. and Skvoretz, J. (2014). A regression analysis of researchers' social network metrics on their citation performance in college of engineering. *Journal of Informetrics* **8**, 667–682.

Costa, B. M. G., da Silva Pedro, E. and de Macedo, G. R. (2013). Scientific collaboration in biotechnology: The case of the northeast region in Brazil. *Scientometrics* **95**, 571–592.

de Arruda, G. F., Peron, T. K. D., de Andrade, M. G., Achcar, J. A. and Rodrigues, F. A. (2013). The influence of network properties on the synchronization of Kuramoto oscillators quantified by a Bayesian regression analysis. *Journal of Statistical Physics* **152**, 519–533. MR3082643

De Stefano, D., Giordano, G. and Vitale, M. P. (2011). Issues in the analysis of co-authorship networks. *Quality and Quantity* **45**, 1091–1107.

Digiampietri, L., Mena-Chalco, J., Silva, G. S., Oliveira, L., Malheiro, A. and Meira, D. (2012). Dinâmica das relações de coautoria nos programas de pós-graduação em computação no Brasil. In *I Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2012)*.

Digiampietri, L. A. and da Silva, E. E. (2011). A framework for social network of researchers analysis. *Iberoamerican Journal of Applied Computing* **1**, 1–24.

Digiampietri, L. A., Mena-Chalco, J. P., Melo, P. O. V., Malheiros, A. P., Meira, D. N. O., Franco, L. F. and Oliveira, L. B. (2014). BraX-ray: An X-ray of the Brazilian computer science graduate programs. *PLoS ONE* **9**, 20. DOI:10.1371/journal.pone.0094541.

Easley, D. and Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge, MA: Cambridge University Press. MR2677125

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks* **1**, 215–239.

Glänzel, W. and Schubert, A. (2005) *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems, Chapter Analysing Scientific Networks Through Co-Authorship*, 257–276. Dordrecht: Springer.

Jackson, M. O. (2008). *Social and Economic Networks*. Princeton, NJ: Princeton University Press. MR2435744

Katz, J. S. and Martin, B. R. (1997). What is research collaboration? *Research Policy* **26**, 1–18.

Krackhardt, D. and Stern, R. (1988). Informal networks and organizational crises: An experimental simulation. *Social Psychology Quarterly* **51**, 123–140.

Latapy, M., Magnien, C. and Vecchio, N. D. (2008). Basic notions for the analysis of large two-mode networks. *Social Networks* **30**, 31–48.

Mählck, P. and Persson, O. (2000). Socio-bibliometric mapping of intra-departmental networks. *Scientometrics* **49**, 81–91. DOI:10.1023/A:1005661208810.

Melin, G. and Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics* **36**, 363–377.

Mena-Chalco, J. P., Digiampietri, L. A. and Cesar Jr., R. M. (2012). Caracterizando as redes de coautoria de currículos Lattes. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2012)*.

Mena-Chalco, J. P. and Cesar Jr., R. M. (2009). scriptLattes: An open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society* **15**, 31–39.

Mena-Chalco, J. P., Digiampietri, L. A., Lopes, F. M. and Cesar, R. M. (2014). Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology* **65**, 1424–1445. MR3444318

Milgram, S. (1967). The small world problem. *Psychology Today* **2**, 60–67.

Nascimento, S. and Beuren, I. M. (2011). Redes sociais na produção científica dos programas de pós-graduação de ciências contábeis do Brasil. *Revista de Administração Contemporânea* **15**, 47–66.

Neal, Z. (2014). The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. *Social Networks* **39**, 84–97.

Newman, M. E. J. (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E, Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* **64**, 016131. MR1975193

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* **69**, 026113. MR2282139

Peron, T. K. D., Costa, L. d. F. and Rodrigues, F. A. (2012). The structure and resilience of financial market networks. *Chaos* **22**, 013117. MR3388494

Said, Y. H., Wegman, E. J. and Sharabati, W. K. (2010). Author-coauthor social networks and emerging scientific subfields. In *Data Analysis and Classification: From Exploration to Confirmation*. *Stud. Classification Data Anal. Knowledge Organ.* 257–268. Berlin: Springer. MR2655162

Senra, N. (2008). Pesquisa histórica das estatísticas: Temas e fontes. *História, Ciências, Saúde* **15**, 411–425.

Senra, N. (2009). Na Primeira República, Bulhões Carvalho legaliza a atividade estatística e a põe na ordem do Estado. *Boletim do Museu Paraense Emilio Goeldi. Ciências Humanas* **4**, 387–399. DOI:10.1590/S1981-81222009000300003.

Stefano, D. D., Fuccella, V., Vitale, M. P. and Zaccarin, S. (2013). The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks* **35**, 370–381.

Travers, J. and Milgram, S. (1969). An experimental study of the small world problem. *Sociometry* **32**, 425–443. DOI:10.2307/2786545.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge, MA: Cambridge University Press.

Yoshikane, F. and Kageura, K. (2004). Comparative analysis of coauthorship networks of different domains: The growth and change of networks. *Scientometrics* **60**, 435–446.

Yousefi-Nooraie, R., Akbari-Kamrani, M., Hanneman, R. A. and Etemadi, A. (2008). Association between co-authorship network and scientific productivity and impact indicators in academic medical research centers: A case study in Iran. *Health Research Policy and Systems* **6**, 1–8.

L. Digiampietri
Escola de Artes, Ciências e Humanidades—EACH
Universidade de São Paulo—USP
Av. Arlindo Bettio
1000 Ermelino Matarazzo
CEP 03828-000 São Paulo
SP Brasil
E-mail: luciano.digiampietri@gmail.com

L. Rêgo
Departamento de Estatística
e Matemática Aplicada
Universidade Federal do Ceará
Av. da Universidade
2853 Benfica
Fortaleza CE 60020-180
CE Brasil
E-mail: leandrochavesrego@gmail.com

F. C. de Souza
Universidade Federal de Pernambuco
Departamento de Ciências Atuáriais
Cidade Universitária
50740-540 Recife
PE Brasil
E-mail: filipecostadesouza@hotmail.com

R. Ospina
Universidade Federal de Pernambuco
Departamento de Estatística
CAST Laboratory
Cidade Universitária
50740-540 Recife
PE Brasil
E-mail: raydonal@de.ufpe.br

J. Mena-Chalco
Centro de Matemática, Computaç
e Cogniç
Universidade Federal do ABC
Av. dos Estados 5001
Bairro Bangu
Santo André
CEP 09210-580
SP Brasil
E-mail: jmenac@gmail.com