Journal of Universal Computer Science, vol. 17, no. 14 (2011), 1951-1970
submitted: 25/3/11, accepted: 28/9/11, appeared: 1/10/11 © J.UCS

Identifying Workgroups in Brazilian Scientific Social Networks

Victor Ströele, Ricardo Silva, Moisés Ferreira de Souza, Carlos Eduardo R. de Mello, Jano M. Souza, Geraldo Zimbrão
(PESC/COPPE - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
{stroele, rick, moises, carlosmello, jano, zimbrao}@cos.ufrj.br)

Jonice Oliveira
(DCC/IM – Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
jonice@dcc.ufrj.br)

Abstract: Social networks are social structures consisting of individuals or organizations, usually represented by nodes tied by one or more types of relationships. Although these structures are often complex, analyzing them enables us to detect several inter and intra connections amongst people in and outside their organizations. In this context, we present an approach using data mining techniques in order to identify intra and inter organizational linkages amongst groups of people with similar profiles. Using clustering techniques, we identify groups of people in a way that allows us to evaluate how researchers collaborate in the Brazilian scientific scenario of Computing Science. Besides this, we are able to understand how research flows amongst the best universities and research centres in Brazil. Understanding the Scientific Brazilian scenario can help the development of research in other scenario or even in other Social Network Types.

Keywords: Data Mining, Group Detection, Scientific Social Networks Analysis, Scientific Collaboration
Categories: J.1 J.4 K.4.2 K.4.3 L.6 L.6.2

1 Introduction

Social networks reflect social structures that can be represented by nodes (individuals or organizations) and their relationships. Relationships can be assigned to specific types of shared interests (such as values, visions, ideas, and religion) or even more specific relationships such as financial exchanges, friendship, communication, conflicts, and others.

Several efforts have been made in order to analyze social networks [Wasserman, 1994] [Freeman, 1979]. From a data mining perspective, the area that analyzes social networks is called link mining or link analysis [Han, 2006].

The purpose of this work was to group people with common characteristics and relationships in a social network and thus provide mechanisms for social networks' analysis. Once grouped, we analyzed the linkages between the people and the groups formed – inside and outside one's company. Based on this analysis we reached some conclusions on the collaboration amongst people and amongst different organizations.

With the aim of evaluating this proposal, we used this approach to study the scientific social network in Brazil, namely in the Computing Science scenario. For this we identified the relationships amongst the best researchers and universities in the computing area in some Brazilian institutions.

Annually, CAPES[1] [CAPES, 2009] -- an institution supported by the Brazilian Government via its Department of Education -- assesses postgraduate programmes and gives them marks or grade scores from 1 to 7. The criteria for this evaluation include: Teaching staff, Research, Formation, and Intellectual Production, amongst other aspects. Institutions rated level seven are considered of the highest excellence. For this work we considered only level seven Computing Science programmes (institutions: COPPE/UFRJ[2] and PUC-RIO[3]) and level six (institutions: UFMG[4], UFPE[5], and UFRGS[6]).

The results obtained through data mining enabled us to identify several social network features. With the analysis of this social network, it was possible to determine the degree of relationship between these educational institutions. Thus, enlarging this approach to analyze the different actors that can be involved in a product design project (researchers, universities, research centres, society, manufacturers, suppliers and so on) is totally possible.

This work is organized in 8 sections, and this is the first one. To follow it, we first present some related works from the literature (section 2) and then explain the scenario we initially applied our solution to: scientific social networks (section 3). In sections 4 and 5 we describe data manipulation and our clustering approach, respectively. Section 6 describes the study case made to evaluate this proposal. Then we conclude our work by pointing out some future steps.

2 Related Works

In this section, we present some related works in group detection, which is one of the challenges found in the link mining area. Group detection is the task of clustering or grouping nodes of a social network that have similar characteristics and are also connected by various relationships with each other.

With the growth of the Web, social networks have recently started to attract the attention of several researchers. A lot of work has been done on the implicit mining communities of Web pages [Gibson,1998] [Flake, 2000] and email [Schwartz, 1993]

[1] CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Higher Education Staff Improvement Coordination.

[2] COPPE/UFRJ – Alberto Luiz Coimbra Institute - Graduate School and Research in Engineering – Federal University of Rio de Janeiro, Brazil.

[3] PUC-Rio – Pontifícia Universidade Católica do Rio de Janeiro – Pontifical Catholic University, Rio de Janeiro, Brazil.

[4] UFMG – Universidade Federal de Minas Gerais – Federal University of Minas Gerais, Brazil.

[5] UFPE – Universidade Federal do Pernambuco – Federal University of Pernambuco, Brazil.

[6] UFRGS – Universidade Federal do Rio Grande do Sul – Federal University of Rio Grande do Sul, Brazil.

[Tyler, 2003]. Other works include Mining newsgroups [Agrawal, 2003] and link prediction [Liben-Nowell, 2003]. Many of the techniques used to identify groups in this scenario can be classified as either agglomerative or divisive clustering methods [Getoor, 2005]. However, the social network needs specific algorithms for clustering, as these algorithms must consider not only the objects' profiles, but also the relationship amongst them.

Clustering algorithms for social networks differ from standard ones, because they must consider the common interests between objects to form groups. In other words, these algorithms must form groups with elements where relationships among them are stronger (most interest in common) while elements with weaker relationships (less interest in common) should be kept in separate groups.

In [Newman, 2004], Newman presents a survey of several clustering algorithms for social networks. These algorithms are based only on the graph structure to identify the groups, i.e., they do not consider the attributes of the nodes. This approach does not focus on finding homogenous groups according to the attributes of the nodes. Therefore, there is no guarantee that the identified groups are homogenous. We say that a group is homogenous when all objects inside the group have a very similar profile. Consequently, finding homogenous groups ensures that elements are similar.

Furthermore, Newman presents different ways to measure the similarity between two nodes. These measures are also based on the network structure, and therefore, it does not take into account the attributes of the nodes and, consequently, does not identify homogenous groups.

The measures, as defined for social networks, can be used in hierarchical clustering algorithms for conventional data, such as the Single Linkage and the Complete Linkage [Han, 2006]. However, the large number of edges that these algorithms must evaluate to identify the groups makes it a time consuming exercise. Still in [Newman, 2004], Newman uses 'edge betweenness', that is, a generalization of 'vertex betweenness' as defined by Freeman [Freeman, 1979] in order to choose edges that will be removed.

In [Tantipathananandh, 2007], Tantipathananandh defined a framework and a set of algorithms to mine social networks that change over time. However, these approaches use only the structure of social networks to find groups. Our approach takes into account both the attributes of individual profiles and the attribute of relationships between each individual.

We analysed several works that examine the social networks formed by relationships as defined by patterns of collaboration [Newman, 2004] [Newman, 2001]. Some of these works examine the social networks formed by relationships of co-authorship [Newman, 2004].

Analysing these related works, we can say that the key difference is that our methodology analyzes both profile attributes and relationship attributes. On the other hand, as said before, the other methods use only the structure of social networks to find groups. Furthermore, our method is faster, as it uses a Minimum Spanning Tree of the graph to build groups, so we do not need to analyse all edges of the graph.

3 Scientific Social Networks

Scientific social networks are social networks where two scientists are considered connected if they have co-authored a paper [Newman, 2001]. The nodes of the graph, that represent the social network that will be examined, are represented by researchers and the edges are relationships between each pair of researchers (see Figure 1).

There are several ways to identify a relationship between two researchers. In general, these relationships may be: Project Participation; Co-authored publications; Advisory work; Examination Board participation; Judgment Committees; Awards; and other types of scientific production (e.g., patents).

The project participation relationship exists only when two or more researchers worked together on a project. Researchers working on the same project have, in addition to performing activities, a common concern about the problem being solved, and so there is a relationship between them.

The relationship of co-authors is one of the most important and most representative items. This is justifiable because researchers are studying and publishing on the same subject. Therefore, there is a common interest between them on the same research subject, so they are more directly related.

The relationship in co-advising occurs when two researchers advise the same student in the same work. So, as well as the relationships of co-authoring, these researchers also develop their research on the same subject or subjects that are related or supplementary.

The advising relationship is the connection between a researcher (mentor) and a student. The identification of this relationship is important for the analysis of relationship evolution over time, as students can become researchers in the future.

The examination board relationship occurs when two researchers participate on the same examination board due to job completion. For example, when two researchers participate on the same examination board for a doctoral thesis presentation, meaning that they have in common, knowledge of what is being presented. Despite it being a weaker link, this is a type of relationship.

The relationships of judgement commissions and awards occur in the same way as those described above, that is, when two researchers participate in the same judgement committees or when they evaluate a possible prize for work, respectively.

The most important relationships in the scientific study of social networks are those that best represent common interests between two researchers. Thus, while all types of relationships are important, the co-authorship relationships are more interesting elements as they represent the interests sought by researchers.

In addition to relationships, each of the researchers has an individual profile, built through one's personal attributes, such as: academic training; research and activity area; number of journal publications; number of proceedings publications; number of technical report publications; number of project participations; number of thesis advising participations; and number of participations on examination boards.

The academic training attribute is the measure of the qualification of a researcher, e.g., M.Sc., D.Sc., / Ph.D., and so on. Research and activity areas indicate what areas of activity the researcher is connected with. Examples of research and activity areas are databases, artificial intelligence, data mining, and software engineering, amongst others. The publication attribute indicates the number of publications a researcher has.

Thus, the publication in journals attribute indicates the number of articles one has published in journals, and so on.

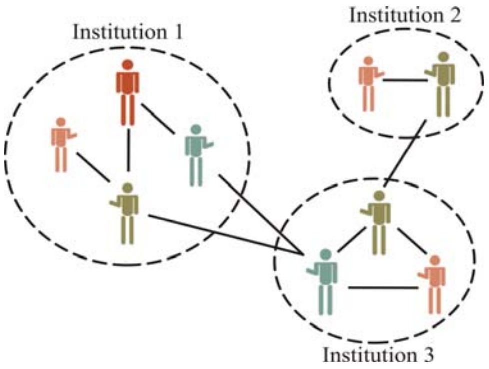


Figure 1: Scientific Social Network Example

Researchers are linked to each other either directly or indirectly. This association may be stronger or weaker according to the degree of the relationship between them. Researchers who have publications in common, working in similar areas and who took part in thesis presentations before, or on examination boards, for example, can be considered as having a strong relationship. On the other hand, if two researchers have participated in only one examination board, the relationship between them is considered weak. There are also cases where researchers are not directly connected, where the relationship will be carried out by other researchers. This data is available on the Lattes Platform[7] [LATTES, 2009].

Each pair of researchers will have a specific degree of relationship between them. The degree of relationship is calculated by adding all common relationships among researchers. Thus, if two researchers have published ten papers together, then they will have relationships with weight/degree ten.

The Lattes Platform is a Web platform established and maintained by the Brazilian Government where researchers and students have to provide information about themselves in a public academic curriculum. It was developed to record the previous and current life of Brazilian scholars, allowing the registration of productive information, such as publications, advising, participation in examination boards, patent registrations, and participation in events, amongst other things. It is interesting that all this data is kept up-to-date by researchers. But the data is only available in the Web. One needs to use a tool to extract it and store it in a database. For that we used the GCC[8] [Oliveira, 2006].

[7] The LATTES Platform is a Brazilian curricular database where researchers and students are registered and publish their research production data. This database is used for many Brazilian production indicators to quantify and measure research production.

[8] GCC is a Web environment originally developed by COPPE/UFRJ whose purpose is to enable knowledge management in research institutions and to improve

After the analysis of the available data we chose the attributes that were consistently met by researchers and that were considered most important in the scientific context. To simplify the description of this work, we present only some of the profile attributes (research and activity areas, journal publications, proceeding publications, and number of theses advised), and only co-authoring relationships.

4 Data Analysis and Pre-Processing

The data used in this work was taken from the researchers' Lattes curriculum of institutions rated level 6 and 7 according to CAPES and as stored in the GCC [Oliveira, 2006]. CAPES' scores go up to 7, that is, the best post-graduate programmes received a score of 7. The analyzed production data goes from 1947 to July, 2007. We had 190 researchers analyzed from five universities who collectively were responsible for 618 project participations; 18,882 co-authored publications; 5,783 advised theses; 4,198 examination board participations; 685 awards; 13 research and activity areas; 2,360 publications in journals; 10,840 publications in proceedings; and 2,248 publications in technical reports.

As mentioned before, we use a data mining technique to analyse data. This technique requires that the data be prepared before it is used. Therefore, we carried out the data pre-processing that consisted of data cleaning, exploratory data analysis, and normalization of selected variables.

In the first pre-processing step, we tried to verify the consistency of the data taken from the GCC database in order to clean the dataset. As a result we found some inconsistencies in the relationships amongst researchers and their bibliographical productions: some researchers were associated to the same production more than once.

Such inconsistencies harmed data analysis and, as a solution, the duplicated relationships were removed from the database. No other inconsistency problems in the data set were detected.

We also analysed the profile and distribution of relationship attributes to identify outliers, and redundancies in variables. To aid the analysis we basically used histograms, box-plots and a correlation matrix. Histograms and box-plots were used to analyze variables distribution, and the correlation matrix was used to analyze the correlation between variables.

4.1 Profile Attributes

During the analysis of researcher profiles, some attributes were removed from the analysis as they allocated many zeros to many researchers. These attributes were: the number of participations in examination board attribute; number of technological productions; number of participations in projects; number of publications in conferences; and number of publications in journals. Part of this problem is caused by researchers who did not fill out information completely in their curricula. The large

collaboration between researchers, stimulating the development of new ideas. Through the services of personnel knowledge management it stores information on the curricula of researchers as obtained from the Lattes Platform.

number of zeros led many researchers to be excluded from the analysis as they were considered outliers in one or more variables. This fact considerably decreased the number of researchers to be analyzed in the group.

Despite the removal of this large number of profile attributes, the influence of the relationships found among researchers helps in the maintenance of cohesive groups. Thus, the reduction of the number of profile attributes did not affect the final result.

During the analysis of the remaining attributes, it was noticed that researchers who have a larger number of classes have a great number of publications. In some cases there are researchers who advise but who do not publish. Therefore, we could not generalize these rules.

CAPES defines the working areas at the moment when a researcher registers his/her curriculum in the platform choosing one or more research areas. However, 45 out of the 190 researchers did not fill out any working area and a great number of registered areas just had an associated researcher. To make the analysis more practical, we choose the areas with more than three researchers and with more representation. Figure 2 shows researcher distribution in these areas.

Figure 2 shows that the areas with greatest interest for research are Software Engineering, Information Systems, and Database. This demonstrates that Brazil has a lack of research in areas such as Hardware, for instance.

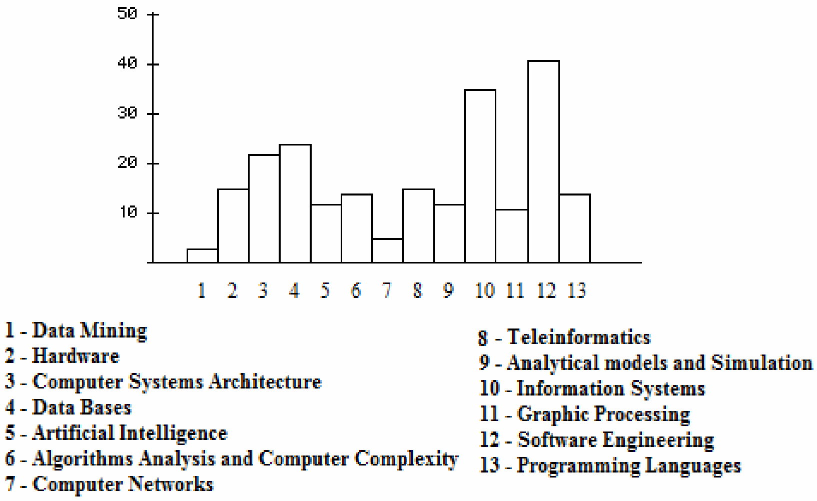


Figure 2: Researcher Distribution in CAPES areas

Having chosen the attributes, we moved onto the data normalization stage. At first we applied the Minimax normalization. However, the amount of zeros obtained with the normalization was high and we opted for the application of natural logarithm $(Y = \ln(X))$ before the application of the Minimax. Figure 3 shows the normal distribution for the bibliographical production attribute without and with the application of the natural logarithm.

The data obtained was analysed to determine possible outliers. Most of the outliers found were related to the bibliographical production attribute: researchers who do not publish or researchers who publish a lot if compared to the average of publications by researcher. The outliers found after the analysis were removed.

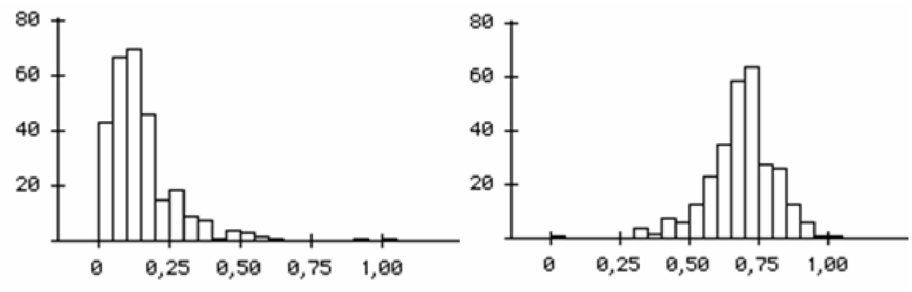


Figure 3: Normal distribution of the bibliographical production attribute without and with the application of the natural logarithm.

4.2 Relationship Attributes

To analyze the relationship attributes, it was necessary to establish a measure that could differentiate weak relationships from strong ones. Initially, the adopted measure was counting the co-authoring productions between researchers. However, most relationships were weak and it was considered a problem: the relationship's weight tended to zero and disappeared when the normalization was applied. The result was a disconnected graph that could not be used as a result of our approach.

In this initial analysis, we found that two researchers had a strong relationship between them but were not related to any other researcher. Thus, we decided to remove both researchers while maintaining all others.

Another problem related to the previous measure is that it did not take the total number of publications by each researcher into account to calculate the relationship degree. In other words, it did not appropriately reflect the relationship degree amongst two researchers in an analysis of the social network.

In order to try an increase of the relationship degree and to have it adequately reflect on it, the formula used to calculate the relationship degree was changed as shown in equation 1:

$$R = \frac{CP}{P1 + P2}. \qquad (1)$$

Where R means relationship degree, CP means number of common publication between researcher 1 and 2, P1 means total publication by researcher 1, and P2 means total publication by researcher 2.

Applying this formula, the data was normalized with the application of the natural logarithm and Min-Max normalization. Although most of the relationships continue to be weak, they have a representative value in relation to the strongest relationships.

With that, we obtained a connected graph as even the weakest relationships continued to be represented.

5 Group Detection Using Minimum Spanning Tree

In this section, we present our proposed group detection method for social networks. This method aims at identifying groups of people in social networks who have similar profiles and strong relationships amongst them. Our method uses a graph approach that reduces the clustering problem to a graph partitioning problem; this method was also used in [Menezes, 2008] and presented in [Mello, 2008].

The proposed method is based on the spatial clustering method described in [Assunção, 2006]. The spatial topology can be understood as a social network of spatial objects where spatial objects are the nodes and their relationships are the edges of the social networks. Thus, we can use the steps of the spatial clustering method to identify the groups in social networks. The main difference between our method and the original method is the modelling of the input graph. The weight edge used in spatial clustering is the dissimilarity measure between attributes of the nodes connected by this edge. In our method, the weight of the edge is defined by the relationship (edge) attributes. This will be clarified when we present the steps of our method. It consists of three steps that we describe in the following subsections.

5.1 Constructing the Social Graph

The data on social networks can be divided into two types: people's profiles and relationships amongst them. The person's profile consists of a vector that stores the person's features or attributes, such as age, weight, height, etc. Features and attributes describe a person. The relationship represents an existing social relationship between two people. The social relationship can be strong between two people and weak between two others. It depends on the chosen measure we use for the relationship.

This step consists of transforming the social network into a graph called social graph. It regards the two types of data presented above. In the social graph, every node represents one person and each edge indicates an existing relationship between two people.

The strength of the social relationship is indicated as the weight of the edges. Therefore, the social graph has weights in all of its edges. The weights of the edges must fall within the interval (0.1), where weights close to 0 indicate a strong relationship and the opposite ones, i.e., weights close to 1, indicate a weak relationship.

These weights can be defined in several ways. For instance, we can use the number of common friends of two persons and suppose that the larger the number of common friends, the stronger the social relationship between them.

In this work, we use the following profile attributes: the number of participations in advising, the number of participations in examination boards, and the number of publications. We use the number of publications in common between each pair of researchers as a relationship. Profile attributes and the relationship have been set in Section 3.

5.2 Minimum Spanning Tree Generation

In this step, we have the social MST constructed. Clustering people in social networks is the same problem as identifying subgraphs in a graph. Considering that partitioning a graph is an NPhard problem, to reduce complexity, we generate the Minimum Spanning Tree (MST) using the weights of the edges.

Minimum Spanning Tree is an interesting structure in the study of social networks because it allows a visualization of the main relationships. An MST has only the strongest relationships of a social network, so the user has a more simplified view of the social graph, enabling him to make some Analyses that would be unfeasible when visualizing all social network relationships.

There are several algorithms for MST generation. PRIM is one of them [Cormen, 2001], and it consists of removing the expensive edges according to weights. In the social graph, the MST represents the strongest social relationships amongst people in a cycle. By now, every edge we prune will form a group of people.

5.3 Pruning the MST

In this step, we remove some edges of the MST to find groups of people with similar profiles. For that, we must calculate the cost of edges related to group homogeneity. Each time an edge is removed we form two groups of people.

The problem is to establish the criteria for selecting the edges that will be eliminated. We define cost with the following formula:

$$\mathrm{cost}(l) = SSD_T - SSD_l, \tag{3}$$

where SSD_T is the sum of the square deviations of the profile's attributes in the T tree to which the l edge belongs. SSD_l is obtained with the following formula:

$$SQD_T = \sum_{j=1}^{m} \sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j \right)^2, \tag{4}$$

where *n* is the number of nodes (people) in *T*, x_{ij} is the j^{th} attribute of the i^{th} person, *m* is the total number of attributes of the profile considered for clustering, and $\overline{x}_j$ is the mean value of the j^{th} attribute amongst all individuals.

The SSD_l part is the sum of the square deviations of the two sub-trees that the *l* edge connects. It is calculated as: $SSD_l = SSD_{T_a} - SSD_{T_b}$, where SSD_{T_a} is the sum of the square deviations of the T_a tree and SSD_{T_b} is the sum of the square deviations of the T_b tree, as shown in Figure 4.

Considering the above, we can say that an edge cost represents a measure of homogeneity. This way, edges with the highest costs are our candidates for pruning. After we prune an edge, we have to re-calculate new costs for all edges of the pruned social graph, as the absence of the removed edge affects calculation results. The edge cost we defined is based on the k-means objective function [Newman, 2001].

Therefore, pruning the highest edge increases homogeneity of the resulting subtrees, i.e., it generates more homogeneous[9] groups of people.

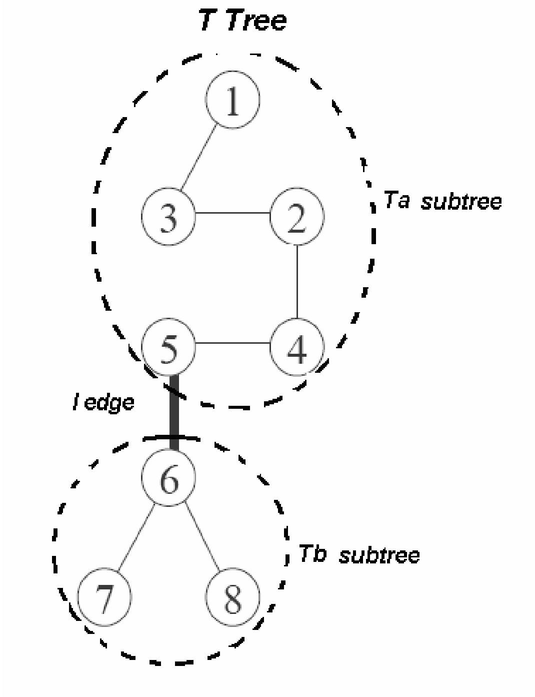


Figure 4:.Cost calculation of the e edge

6 Results

The case study of this work aims at identifying research communities within and amongst Brazilian universities, using the data described in section 4. In this case study, we only use Brazilian universities and the domain of Computing Science. Altogether, we have five universities and 190 researchers.

We use three attributes of the researchers' profiles to identify the groups. We consider the more important ones: the number of publications in journals, the number of papers in proceedings, and the number of holdings in guidelines. We also use the co-authorship as attributes of relationship.

Figure 5 shows the Minimum Spanning Tree generated by the PRIM algorithm [Cormen, 2001] as a result of the proposed methodology. The largest colored regions illustrate the Brazilian institutions, and the smaller colored boxes with the same number illustrate a different group within an institution. In Figure 5 each number represents a cluster generated by the group detection method. The edges represent

[9] By homogeneous groups, we mean groups of people who are similar in terms of characteristics (i.e., working in the same area).

only the strongest relationships. The colours were used to help with visualization of the regions and groups.

Analyzing Figure 5 we have a global view of the Scientific Social Network. This view is used in evaluation of the first results and allows us to understand the main relationships among institutions, groups and also between researchers.

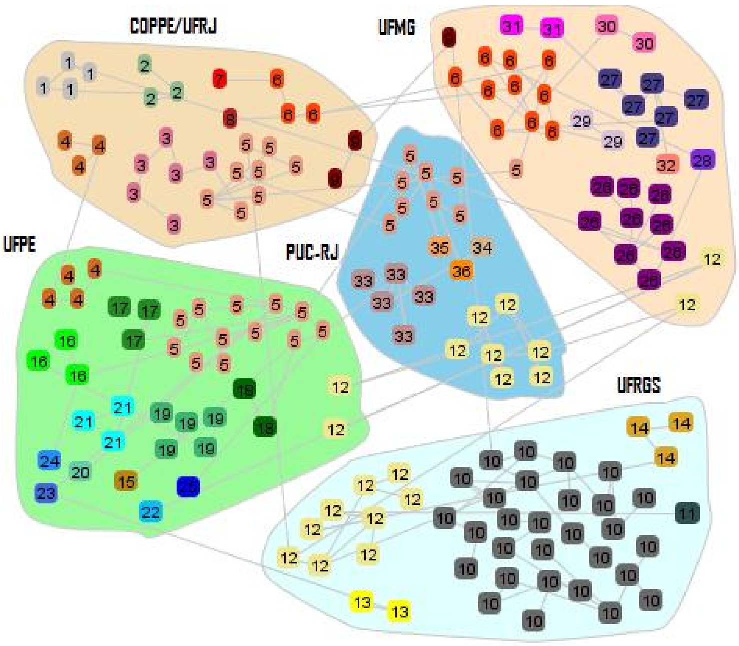


Figure 5: Inter and Intra institutional relationship

6.1 Clustering Analysis

In this session we analyze the groups identified. First, we look at the groups within a specific university. After that, we analyze groups generated amongst universities. Finally, we validate the groups.

Table 1 shows group distribution within and amongst institutions. The diagonal of this table represents the number of groups for a specific university, i.e., cells ii represent the number of groups that exists only in university i.

Analyzing the results from Table 1, we can see that most groups belong to a particular university. But we can also see that there are some groups that belong to two or more universities, i.e., there is cooperation amongst institutions.

	UFRJ	PUC-RJ	UFMG	UFRGS	UFPE
UFRJ	5	1	3	-	2
PUC-RJ	1	4	2	1	2
UFMG	3	2	7	1	2
UFRGS	-	1	1	4	1
UFPE	2	2	2	1	11
TOTAL	11	10	15	7	18

Table 1: Group distribution within and amongst universities

Looking at Table 1, we can see that UFRJ and UFMG are the universities that have a greater number of groups in common. UFMG is the institution with the largest collaborative effort with other institutions, because it has the largest number of groups in common with other universities; UFRGS, however, is the institution with the lowest collaborative effort, with the least number of groups in common.

Analysing Table 1, we can see that UFRGS and PUC-RJ are the ones that have a more uniform collaboration amongst universities. They have one or two groups collaborating with other institutions and have only four internal groups. On the other hand, UFPE and UFMG have many internal groups (eleven and seven, respectively) and a few inter-institutional groups.

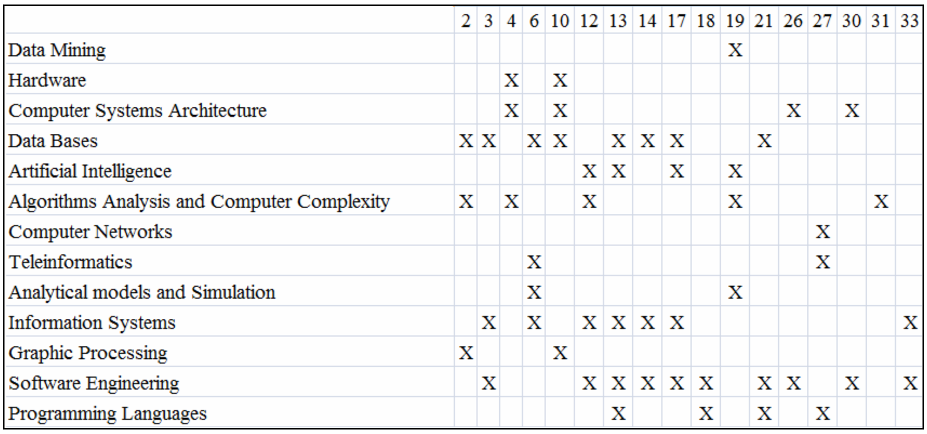
Table 2 shows group distribution within the universities. This table shows that 14 of the 31 internal groups have only one individual. The groups that have only one researcher are called “Unit Groups”. We must remember that these researchers do not work alone; they must have at least one relationship with a different researcher, but this relationship is not strong. This may be related to beginner researchers, researchers who usually publish only with their students, or even researchers with few publications.

	Unit Groups	Internal Groups
UFRJ	2	5
PUC-RJ	3	4
UFMG	2	7
UFRGS	1	4
UFPE	6	11
TOTAL	14	31

Table 2: Group distribution within the universities

We can see in Table 2 that most of the researchers in unit groups are at UFPE. UFRGS is one of the universities that has least cooperation with the others - only three groups (Table 1), and is the institution with the least number of unit groups (Table 2).

We use the researchers’ areas of expertise to validate the groups. We expected researchers within the same group to do their work in the same area, or in similar areas. Table *3* shows the areas and the groups analyzed. As mentioned above, not all areas were analyzed. The unit groups and the groups that have researchers without expertise in these areas were not analyzed. Table *3* shows that the groups really have similar expertise areas.



	2	3	4	6	10	12	13	14	17	18	19	21	26	27	30	31	33
Data Mining											X						
Hardware			X		X												
Computer Systems Architecture			X		X								X		X		
Data Bases	X	X		X	X		X	X	X			X					
Artificial Intelligence						X	X		X		X						
Algorithms Analysis and Computer Complexity	X		X			X					X					X	
Computer Networks														X			
Teleinformatics				X										X			
Analytical models and Simulation				X							X						
Information Systems		X		X		X	X	X	X								X
Graphic Processing	X				X												
Software Engineering		X				X	X	X	X	X		X	X		X		X
Programming Languages							X			X		X		X			

Table 3: *Expertise areas for each group*

6.2 Relationship Analysis

The above analysis shows that the groups found after the use of the method presented in Section 6 are consistent. Starting from this conclusion, several analyses considering the relationships' point of view in the social network will be presented.

The manner in which results are presented allows us to analyze the social network globally, locally, and in a localized way. Through a global perspective it is possible to see all the relationships between the educational institutions. The local analyses evaluated the relationships between different research areas with the aim of identifying areas of common interest. Finally, a localized analysis examines the relationships of a special research member.

Initially, the conclusions derived from the global view of social networks will be presented. According to it, a study that considered the number and strength of the relationships between the educational institutions was done. For a better understanding of the following analysis, the relationships were classified into two types: internal and external. The relationships between researchers who belong to a single institution are called internal, whereas external relationships are those that connect two researchers belonging to different institutions.

	UFRJ	**PUC-RJ**	**UFMG**	**UFRGS**	**UFPE**
UFRJ	–	126	101	164	124
PUC-RJ	126	–	132	226	175
UFMG	101	132	–	169	128
UFRGS	164	226	169	–	221
UFPE	124	175	128	221	–
TOTAL	**515**	659	530	**780**	648

Table 4: Total number of relationships amongst institutions

We examined the number of relationships between each institution, building a symmetric matrix, as shown in Table 4, where pair ij represents the total relationships between institution i and institution j, for i≠j. The sum of the columns of each institution represents the total external relationships for that institution. So, it can be seen that UFRGS is the institution that has the most work with other institutions. On the other hand, UFRJ is the university that has the fewest researchers linking up with other institutions.

The analysis of the strength of the relationships between institutions followed the same steps as the study presented earlier. Table 5 also shows a symmetric matrix, where the pair ij represents the total of strong relationships between institutions i and j, so that we can identify the institutions that are more closely related. The strong relationships were calculated with the aid of the minimum spanning tree -- built by the method presented in this work, as shown in Figure 5. Each relationship indicates that a researcher from institution i has a large number of publications with researcher j, that is, there is strong cooperation between these two scholars. Looking at Table 5, it is possible to see that UFMG is the institution that is most strongly linked to other institutions, while UFRGS is less strongly linked to the others.

	UFRJ	**PUC-RJ**	**UFMG**	**UFRGS**	**UFPE**
UFRJ	–	1	5	1	1
PUC-RJ	1	–	1	0	3
UFMG	5	1	–	2	3
UFRGS	1	0	2	–	1
UFPE	1	3	3	1	–
TOTAL	8	5	**11**	**4**	8

Table 5: Number of strong relationships amongst institutions

Comparing the two results presented earlier, it was concluded that there are some researchers who have the profile of working more closely with another researcher from another university. Thus, the relationship between the institutions is linked to a small group of researchers. This fact can be seen through the relationships between UFRJ and UFMG. On the other hand, there are universities where their external relationships are formed by a large number of researchers. Therefore, this type of institution has several researchers with weak external relationships and not just some researchers with strong relationships. This can be seen in UFRGS, which is the institution with the largest number of external relationships, but, on the other hand, the lowest number of strong relationships.

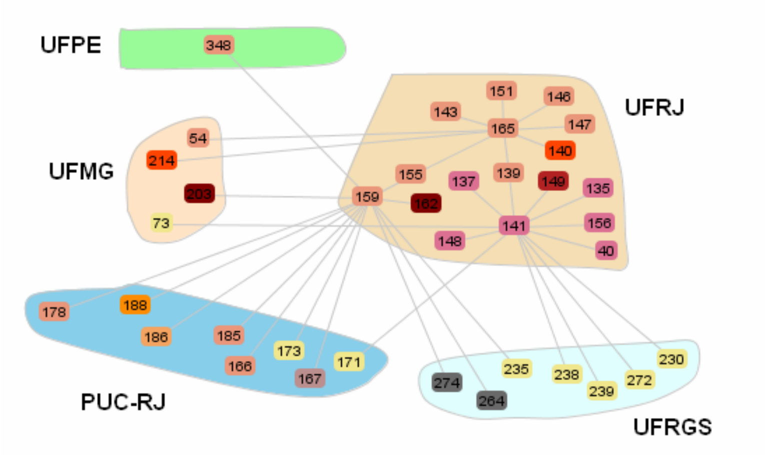


Figure 6: Local view of the social network

The degree of relationship between researchers was evaluated considering the social network in a more detailed way. This analysis was based on internal and external relationships as defined above, from the point of view of each researcher. As expected, internal relationships are generally stronger than external ones. Thus, it was found that researchers from the same institution have a greater tendency to publish together than researchers from different institutions.

In the attempt to identify other characteristics of the social network, the relationships of some specific researchers were examined. The idea was to identify the critical points of the social network, or to identify the researchers that are central to knowledge. In that context, we could see that researchers had different relationship profiles. In most cases, they had a profile for internal relationships, a profile for external relationships, or a profile for both internal and external relationships.

To examine these relationships we selected three researchers who have these relationship types and built a social network that only presents data from researchers 141, 159 and 165. Figure 6 shows a local view of the social network that facilitates the viewing and analyzing of researchers who have the relationship profiles described above. In this picture the numbers represent each researcher of the scientific social network.

Researchers with an internal profile are those with more internal relationships than external relationships. Figure 6 show that researcher 165 is a knowledge centralizer and has a profile for internal relationships.

On the other hand, external researchers, who are a minority in the social network of educational institutions, are those with more external than internal relationships. We can say that this kind of researcher is a minority because it is rare we find researchers who are more connected to external institutions than with their own university. Looking again at Figure 6, it is possible to see that researcher 159 has only two internal relationships and twelve that are external. Thus, researcher 159 is a great collaborator with other institutions, but has a very little internal collaboration.

Finally, there is still a group of researchers who have internal and external relationships in the same proportion. The latter case is illustrated by researcher 141 who, besides being an internal centralizer, also has several external relationships. An interesting point about this researcher is that all of his external relationships are conducted with researchers who belong to the same group. Thus, it is possible to conclude that these researchers have an external line of research similar to that of researcher 141.

7 Result Validation

All the data used in the experiments described above is official: they are derived from researchers curriculum lattes. As shown in the previous section, the data represents the academic profile of each researcher. Thus, as the data is real, we have guarantees on the validity of the structure of the social network that was studied. In addition to data consistency, we confirm the validity of the analysis of relationships and groups formed by the method proposed using a qualitative evaluation. We interviewed – with the support of a questionnaire - the researchers of one of the universities, analysed the answers and then compared them with the results of our approach. In this questionnaire we included questions that would allow us to identify some features, such as: if the areas are interlinked; how the relationship amongst the researchers is; and, consequently, how the relationship amongst the different institutions is. Also, researchers were asked about their areas of interest: if he/she works with researchers from other areas; if he/she is usually a co-author with researchers from other institutions; and other kinds of questions, to map the collaborative scientific production. In addition, we requested researchers to indicate how often they write with internal researchers and how often with external researchers. Thus, it was possible to determine if the researcher has an internal or external relationship profile. We also tried to identify the researchers with whom one had had more publications. Thus, we found key relationships of the researchers and consequently identified their strongest relationships. Initially this questionnaire was distributed only to COPPE/UFRJ researchers.

During this qualitative evaluation, the researchers indicated areas in which they published the most. Thus, we identified the interdisciplinary areas, and, in some cases, we saw that relationships amongst researchers from different areas are very strong. Hence, these researchers are members of the same research group, as we identified by using our approach -- validating the interdisciplinary groups formed by the group detection method.

We saw that the areas focused on -- databases, software engineering and information systems -- are strongly related. Thus, we validated the issue that there are groups formed by researchers who belong to these areas. The same case occurs, for example, with the areas of artificial intelligence and information systems.

We also proved the degree of existing relationships. The majority of the researchers who answered the questionnaire say they have more publications with researchers in the same institution than with external researchers. Each researcher indicated the name of the researcher who is most related to him. We could then validate the strong and weak relationships as well as critical points in the social network.

External relationships (COPPE/UFRJ with other universities) were also proven. The results obtained with the questionnaire show that both the method proposed as well as the analyses of the relationships are correct.

8 Conclusion and Future Work

The paper presents an interesting analysis of scientific collaborations within the Brazilian research community using social network analysis. The Brazilian case serves to illustrate the possibilities of using social network analysis to study research collaborations in general.

In this paper we used a group detection method to identify research communities in the Brazilian scientific social network that could be used in other contexts (with a few small changes). With the use of this method, we obtained a result that allowed us to make a detailed analysis of the social network. We analyzed both the groups and the relationships amongst researchers.

We therefore looked at various aspects of the social network, such as: identification of interdisciplinary areas; level of cooperation between institutions; strong and weak relationships; researchers who play a centralizing role in the social network, etc. With the analysis presented in this article, it is possible to identify evolution points in scientific collaboration both between and among researchers and educational institutions and, based on such analysis, to identify ways of improving scientific collaboration in Brazil.

One of the future works is to expand these qualitative analyses to other universities, to take a complete validation of our approach.

We will also use different weight indicators (threshold) to analyze inter and intra relationships. We will also review the other Brazilian Computing Science institutions -- level 3, 4 and 5. Thus, we will have a wider analysis of results, taking into account all Computing Science post-graduate programmes in Brazil. In this context, we want to try to improve the collaboration level of institutions and researchers according to their preferences, grouping them, and making recommendations. Like Jung [Jung, 2005], this is done in a process of three tasks: i) collecting relevant feedback, ii) grouping like-minded users, and iii) propagating recommendation.

This approach is being developed in conjunction with the work aimed at balancing the social networks [Monclar, 2007] and will be a powerful tool in the analysis of social networks of educational and scientific institutions, as well as other organizations.

Regarding other future work, we want to implement a computer tool to analyze the collaborative design of the scientific social network analyzed and improve it through the suggestion of new relationships. This tool will also be used to suggest new workgroups or research teams to better propagate the knowledge in the social network, thus improving scientific and technological innovation.

Acknowledgements

This research was funded by CAPES, CNPq, and by the COPPETEC Foundation.

References

[Agrawal, 2003] Agrawal, R., Rajagopalan, S., Srikant, R., Xu, Y.: Mining newsgroups using networks arising from social behavior. In Proceedings of 12th International World Wide Web Conference, 2003.

[Assunção, 2006] Assunção, R.M., Neves, M.C., Câmara, G., Freitas, C.C.: Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees, In International Journal of Geographical Information Science, v. 20, n. 7, pages 797–811, August, 2006.

[Axelrod, 1984] Axelrod, R.: The Evolution of Cooperation, New York/ Basic Books (1984).

[CAPES, 2009] CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, 2009, http://www.capes.gov.br.

[Cormen, 2001] Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 2nd Ed., USA, MIT Press and McGraw-Hill, 2001.

[Flake, 2000] Flake, G.W., Lawrence, S., Giles, C.L.: Efficient identification of Web communities. In Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD-2000), 2000.

[Freeman, 1979] Freeman, L.: Centrality in social networks: Conceptual clarifications. Social Networks, 1:215-239, 1979.

[Getoor, 2005] Getoor, L., Diehl, C.P.: Link mining: A survey. SIGKDD Explorations, 2(7), 2005.

[Gibson,1998] Gibson, D., Kleinberg, J., P. Raghavan.: Inferring Web communities from link topology. In Proceedings of the 9th ACM Conference on Hypertext and Hypermedia, 1998.

[Han, 2006] Han, J., Lamber, M.: Data Mining: Concepts and techniques, 2nd Ed., USA, Morgan Kaufmann Publishers. 2006.

[Holme, 2002] Holme, P., Kim, B.J., Yoon, C.N., Han, S.K.: Attack vulnerability of complex networks, Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top, 65, 056109, 2002.

[Jung, 2005] Jung, J.J.: Visualizing Recommendation Flow on Social Network, Journal of Universal Computer Science (J.UCS), vol. 11, No. 11 (2005), 1780-1791.

[Klein, 2001] Klein, M., Sayama, H., Faratin, P., Bar-Yam, Y.: What complex systems research can teach us about collaborative design, Proceedings of the Sixth International Conference on Computer Supported Cooperative Work in Design, London, Ontario, Canada, July 12-14, 2001, pp. 5-12.

[LATTES, 2009] LATTES, http://lattes.cnpq.br/.

[Li, 2008] Li, W. and Shenb, W.: Collaborative design: New methodologies and technologies, S.I. Advances in Collaborative Engineering: From Concurrent - SI on Collaborative design: New methodologies and technologies - Editorial. December 2008, pp. 853-854.

[Liben-Nowell, 2003] Liben-Nowell, D., Kleinberg, J.: The Link Prediction problem for social networks, In Proceedings of the 12th International Conference on Information and Knowledge Management, pages 556-559, 2003

[Mello, 2008] Mello, C.E.R.: Agrupamento de regiões: Uma abordagem utilizando acessibilidade, M.Sc. Thesis, Rio de Janeiro, Brazil, 2008.

[Menezes, 2008] Menezes, V.S.A., Silva, R. T., Souza, M. F., Oliveira, J., Mello, C. E. R., Souza, J. M., Zimbrao, G.: Mining and Analyzing Organizational Social Networks Using Minimum Spanning Tree, Proceedings of 16th International Conference Cooperative Information Systems (COOPIS 2008), 2008.

[Monclar, 2007] Monclar, R.S., Oliveira, J. ; Souza, J. M.: A New Approach to Balance Social Networks. In: Proceedings of UK Social Network Conference, 2007. p. 141-142.

[Mueller-Prothmann, 2004] Mueller-Prothmann, T., Finke, I.: SELaKT - Social Network Analysis as a Method for Expert Localisation and Sustainable Knowledge Transfer, Journal of Universal Computer Science (J.UCS), vol. 10, no. 6 (2004), 691-701

[Newman, 2001] Newman, M.E.J.: The structure of scientific collaboration networks, Proceedings of the National Academy of Science USA 98, 404-409, 2001.

[Newman, 2004] Newman, M.E.J.: Detecting community structure in networks. European Physical Journal B, 38:321-330, 2004.

[Oliveira, 2006] Oliveira, J., Souza, J.M.D., Miranda, R., Rodrigues, S., Kawamura, V., Martino, R.N.D., Mello, C.E.R.D., Krejci, D., Barbosa, C.E., Maia, L.: "GCC: A Knowledge Management Environment for Research Centers and Universities". 8th Asia-Pacific Web Conference, 2006.

[Preece, 2004] Preece, J.: Etiquette, Empathy and Trust in Communities of Practice: Stepping-Stones to Social Capital, Journal of Universal Computer Science (J.UCS), vol. 10, no. 3 (2004), 294-302

[Schwartz, 1993] Schwartz, M.F., Wood, D.C.M.: Discovering shared interests using graph analysis. Communications of the ACM, 36(8):78–89, 1993.

[Tantipathananandh, 2007] Tantipathananandh, C., Berger-Wolf, T., Kempe, D.: A Framework for community Identification in Dynamic Social Networks. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 717-726, 2007

[Tyler, 2003] Tyler, J.R., Wilkinson, D.M., Huberman, B.A.: Email as Spectroscopy: Automated Discovery of community Structure Within Organizations. In Proceedings of the First International Conference on Communities and Technologies, edited by M. Huysman,E. Wenger, V. Wulf (Kluwer, Dordrecht, 2003)

[Wasserman, 1994] Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge, UK, 1994.