*Journal of Universal Computer Science, vol. 18, no. 8 (2012), 1048-1068*
*submitted: 22/9/11, accepted: 14/12/11, appeared: 28/4/12 © J.UCS*

# Modeling, Mining and Analysis of Multi-Relational Scientific Social Network

**Victor Ströele, Geraldo Zimbrão, Jano M. Souza**
(COPPE/UFRJ – Graduate School of Computer Science
Federal University of Rio de Janeiro, Brazil
{stroele, zimbrao, jano}@cos.ufrj.br)

**Abstract:** Social networks are dynamic social structures consisting of individuals or organizations, usually represented by nodes tied by one or more relationship type. Analyzing these structures enables us to detect several inter and intra connections between people in and outside their organizations. In this context, we construct a multi-relational scientific social network where researchers may have four different types of relationships with each other. We adopt some criteria such as relationship age in order to assign a weight to relationships and to enable the modeling of a scientific social network as close as possible to reality. Using clustering techniques with maximum flow measure, we identify the social structure and research communities in a way that allows us to evaluate the knowledge flow in the Brazilian scientific community.

**Keywords:** Knowledge Flow, Multi-relational Scientific Social Network Analysis, Max Flow Grouping Algorithm, Weighted Relationships
**Categories:** J.1, J.4, K.4.2, K.4.3, L.1.0, L.6.0, L.6.2

## 1 Introduction

### 1.1 The problem

Nowadays, there are lots of data on the Web with social characteristics, for example, personal pages, collaborative and social tools; and also, data used in scientific scenarios, such as personal curricula, raw data, publications and articles, books, patents, web services providing access to scientific models, and other kinds of data. Many of these data are used as sources of research by researchers for the development of their scientific work. Currently, the Web is the main interface used by researchers to search for works related to what is being developed by them. Often a researcher uses the Web to find articles, books, or even another researcher that dominates the subject to help in the development of his work. Thus, researchers create links to each other by direct or indirect collaboration, citation and work evaluation.

This phenomenon is allowing researchers to study how connections between people are established and how they evolve over time. Several efforts have been made to analyze social networks in order to help the study of social structures [Wasserman, 1994] [Freeman, 1979]. The connection and the people are represented as social networks.

A network is a set of objects where each one of them is connected to another one. We can represent a network as a graph where the nodes, or vertices, are related or not

related, by edges. A social network reflects a social structure that can be represented by individuals or organizations and their relations. In general the relations represent one or more types of interdependency (such as idea and religion) or more specific relationships (like knowledge/information exchanges and friendship). Thus, with this social structure we can study the exchange of data and information between the individuals or organizations.

In the real world, social networks are mostly multi-relational, i.e., persons or institutions are related through different relationship types. One of the proposals presented in this work is to build a multi-relational social network which allows developed studies to consider the influence of all relationships to each individual.

Based on a multi-relational scientific social network we look for research communities, i.e., we would like to identify groups of researchers who have common interests in developing their research. We use a clustering technique to find research communities and, moreover, we analyze in detail the identified communities and the social network as a whole. Finally, it was possible to define the flow of scientific information in social networks within Brazil.

### 1.2 Related Work

The growth of social networks is due to the evolution of the Web. This growth has attracted the attention of many researchers who intend to analyze these social networks. A lot of work has been done at mining communities of Web pages [Gibson, 1998] [Kumar, 1999] [Flake, 2000] and e-mail [Schwartz, 1993] [Bird, 2006] [Tyler, 2003]. Other works include Mining newsgroups [Agrawal, 2003] and links prediction [Liben-Nowell, 2003]; discover useful information and patterns from data streams on sensor networks [Jung, 2010(a)][Jung, 2011].

The first stage of our work was to study clustering algorithms. Newman [Newman, 2004(a)], Han and Kamber [Han, 2006] examined various clustering algorithms in networks. They concluded that there are still many things to be developed, new algorithms to be constructed, and improvements to be made to existing algorithms.

The second stage of our work was to construct and analyze our multi-relational social network. There are many types of applications based on social network analysis, such as dark networks [Raab, 2003] [Pioch, 2005], content-based recommendation systems [Huang, 2005] [Golbeck, 2005], link predictions [Liben-Nowell, 2003] [Farrel, 2005] [Lim, 2005] [Zhu, 2003], Economic Networks with cooperative and non-cooperative interaction [Jackson, 2010], mobile recommendation service [Jung, 2010(b); Jung. 2012], among others.

In this second stage we constructed and analyzed scientific relationships of Brazilian researchers. We analyzed several works that examine the social networks formed by relationships defined by patterns of collaboration [Newman, 2000] [Newman, 2001(a)] [Newman, 2004(a)]. Some of these works examine the social networks formed by relationships of co-authorship [Newman, 2000] [Newman, 2001(b)] [Newman, 2001(c)] [Newman, 2004(b)] [Ichise, 2005][Jung, 2010(c)].

In addition to examining how researchers exchange knowledge, we hope to identify knowledge brokers, i.e., researchers that are extremely important in maintaining the knowledge flow of the network [Cross, 2004].

For many scientific social network problems only co-authorship is analyzed, but it is not difficult to see that social networks can have many other types of scientific relationships. Although the analysis of multi-relational social networks is quite interesting, few studies have been done in this area [Stroele, 2009] [Cai, 2005] [Jung, 2007].

In this paper we analyze the knowledge flow in a multi-relational scientific social network. To this end, we use the clustering algorithm k-medoids to find research communities, and make a detailed analysis of intra and inter-institution relationships.

### 1.3 Our Contributions

From a data mining standpoint, a social network is a multi-relational data set represented by a graph [Han, 2006]. The graph is typically very large, with nodes corresponding to objects and edges to relations between objects. In the real world the edges usually represent different types of relationships. In data mining, the area that studies social networks is called link mining or link analysis [Liben-Nowell, 2003] [Han, 2006] and one of the challenges for link analysis is group detection, which is the identifying of groups of objects that belong to the same group or cluster.

This paper shows the use and evaluation of our approach in the identification of scientific relationships, based on researchers' curricula data available on the Web. Even though the study of multi-relational scientific social networks identifies researchers with the same research objective, we can also identify cross-cutting groups. These groups are built by researchers belonging to different research areas or laboratories in an institution. Identifying these groups helps in the discovery of interdisciplinary groups.

In addition, one of the themes explored is that of knowledge brokers: people who, by virtue of their relationships with people in different organizations serve as boundary spanners (moving information and context from one group to another) or bottlenecks (impeding the flow of information and context) [Jackson, 2010]. As the relationships of our social network involve a collaborative concept, we will identify knowledge brokers who are boundary spanners.

In general, studying the formation of these networks allows us to identify how researchers and organizations are working. We can find the degree of involvement between researchers, research areas and organizations. We can also identify patterns in cooperation (inter and intra-university). All these possibilities can help us to improve the Web of Data through the promotion or provision of greater collaboration between researchers, resulting in more quality data and more publications.

Scientific social networks allow interaction between experts from different areas. Analyzing these networks leads to improvement in collaboration and knowledge flow. This impacts on the quality of scientific production, improving it through more use, evaluation, recommendation and collaborative filtering, by the scientific community.

With the purpose of identifying the points raised above, the goal of this work is to group people with common relationships in the scientific social network, by means of data mining techniques. We use this approach to study the multi-relational scientific social network in Brazil. We analyze the links and reach some conclusions on the collaboration among people and among organizations. The results obtained enable us to identify several features of the multirelational scientific social network.

Section 2 presents multi-relational social networks. Section 3 describes how to model a multi-relational network. Section 4 explains the data mining methodology. Section 5 shows the results and analyses of the scientific social network and the result validation. Finally, in Sections 6, we focus on conclusions and future works.

## 2 Multi-relational Scientific Social Network

There are two types of social networks: Homogeneous and Heterogeneous [Cai, 2005]. Homogeneous Social Networks are those where there is only one kind of relationship between objects (Figure 1(a)). Heterogeneous Social Networks represent several kinds of relationships between objects and are also known as Multi-relational Social Networks (Figure 1 (b)).

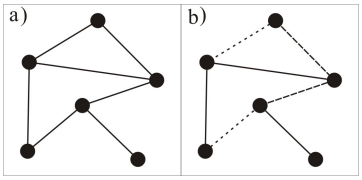


*Figure 1: a) Homogeneous Social Network; b) Multi-relational Social Network.*

Indeed, most social network mining methods consider only Homogeneous Social Networks. However, in the real world, almost all social networks have several kinds of relationships between objects.

A particular kind of problem concerning Multi-relational Social Networks lies in extracting the different relations that exist within them. Each relation can be modeled as a graph. Depending on the information that you want to obtain, analyzing one of the relations will be more important than the others. Thus, for a better analysis of the multi-relational social network you need to select the relations that have a positive effect on one's proposal [Cai, 2005].

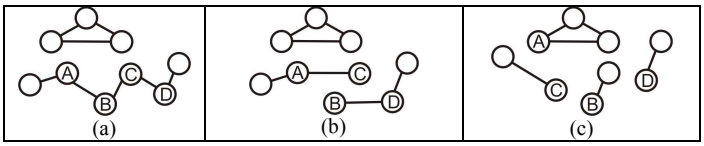


*Figure 2: Multirelational Social Network with three types of relationships.*

Figure 2 is an example of a social network with three different types of relationships. Depending on the information you want to attain it will be necessary to analyze the social network from the point of view of a particular relationship. Thus, if a user wants objects A, B, C and D to belong to the same community, then the relationship depicted in Figure 2(a) will have a positive effect on the information, whereas the relationship depicted in Figure 2(c) has a negative effect on the

analysis. Consequently, the user has to know what he wants to discover in the social network, and then define the relationship to be used.

Scientific social networks are social networks where two scientists are considered connected if they have co-authored a paper [Newman, 2000] [Newman, 2001(a)] [Newman, 2001(b)] [Newman, 2001(c)] [Ichise, 2005]. These networks are more complex as their relationships involve different types of scientific collaboration or interaction. Thus, we can consider that Scientific Social Networks are a kind of Multi-relational Social Network.

In Homogeneous social networks, where there is only one type of relationship, the knowledge flow is through this specific relationship. Thus, Social Network Analysis considers only one type of knowledge exchange between network elements.

On the other hand, in multi-relational networks the knowledge exchange takes place through different kinds of relationships. Thus, multi-relational social network analysis assumes that elements are exchanging different knowledge types depending on the types of relationships linking them.

Analyzing different relationship types also allows elements related by secondary relationships to have their connections reflected in social networks. For example, two researchers who have no co-authorship in common but who are both part of the same project will have an explicit relationship in a multi-relational scientific social network.

In our multi-relational network we consider that researchers have stronger or weaker links according to the degree of the relationship between them. For example, researchers who have publications in common, work in similar areas and have taken part in thesis presentations before can be considered to have a strong relationship. Whereas, if two researchers have participated in only one examination board then the relationship between them is considered weak.

In this work, we use four different relationship types to model our Multi-relational Scientific Social Network: Project Participation; Co-authored publications; Advisory work; and Technical Production.

The *Co-authored Relationship* is one of the most important items. This is due to the fact that researchers are studying the same subject. Therefore, there is a common interest between them on the same research subject, so they are more directly related.

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

The *Project Participation* occurs when two researchers work together in the development of the same scientific project.

The *Technical Production* exists when two or more researchers developing together some technical works, software development, and so on.

The next section shows how we constructed our Weighted Multi-relational Social Network, i.e., we modeled a Social Network where each link carries a different weight, representing how close two researchers are to one another.

## 3 Multi-relational Model

The data used in this work is available on the Lattes platform [LATTES, 2006]. This is a Web platform set maintained by the Brazilian Government where researchers and

students provide their information in a public academic curriculum. To extract and store this data in a database we used the GCC[1] [GCC, 2006].

This data was taken from the researchers' Lattes curriculum of institutions rated levels 6 and 7 according to CAPES (scores go up to 7). CAPES is the Federal Post-Graduate Staff Improvement Coordination which plays a key role in the expansion and consolidation of MSc and PhD courses in Brazil. In addition, this institution is responsible for evaluating these post-graduate courses. Altogether we analyzed 175 researchers from the Computer Science area from five Brazilian universities.

In modeling a social network relationship the weight represents how strongly two elements are connected. The weight of a link in a multi-relational social network should consider the weight of all relationship types between two elements.

The process of multi-relational network modeling was divided into three steps: number of common relationships between researchers; age of relationships that link these researchers; and loss of knowledge when the relationship between researchers is indirect.

### 3.1 Number of Relationships in Common

To analyze the relationship attributes it was necessary to establish a measure that could differentiate weaker relationships from stronger ones. The first action was to count the number of project participations, co-authored publications, advisory work, and technical productions between researchers. Researchers with the largest number of interactions have a stronger relationship.

In the data mining methodology adopted for this work, the weight of a relationship should be in the interval [0, 1], so these values were normalized. In order to try to increase the relationship degree we used the formula represented in equation (1):

$$R_i = \frac{CR_i}{P1 + P2}, \quad i = 1, ..., t. \tag{1}$$

Where $R_i$ means the degree of relationship 'i', $CR_i$ is the number of common relationships 'i' between researchers 1 and 2, P1 and P2 means total publications by researcher 1 and 2, respectively, and 't' is the total of relationship types.

We divide the number of common relationships by the sum of total relationships of each researcher so that the relationship strength is relative to the total number of relations. Thus, we prevent relationships with the same frequency from having the same strength.

After applying this formula to the four types of relationships we sum all the degrees of these relationships ( $R_i$ ) and assign weights to each of them. Equation (2) represents this step

$$TR_{AB} = \sum_{i=1}^{t} \alpha_i R_i, \tag{2}$$

[1] GCC is a Web environment originally developed by COPPE/UFRJ whose purpose is to enable knowledge management in research institutions and to improve collaboration between researchers, stimulating the development of new ideas.

where $TR_{AB}$ means the total number of relationships between researchers $A$ and $B$, $\alpha_i$ is the weight of relationship 'i'. In this work we have used $\alpha = 1$ for all types of relationships. However, we modeled the multi-relational scientific social network so that it can be used in any problem.

In order to illustrate the concept presented earlier for relationships with different weights, consider the social network formed by companies and their relationships, illustrated in Figure 3. In this network we have six companies linked by three relationship types: competition, partnership cooperation, and financial interest collaboration.

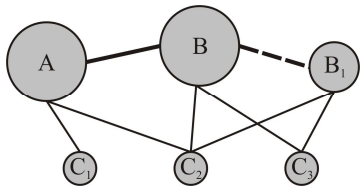


*Figure 3: Collaborative and non-collaborative relationships*

The *competition relationship*, represented by the relationship between companies A and B, are existing relationships among companies that compete for the provision of common services. The *partnership collaborative relationship* occurs between B and B1, shown in Figure 3. This type of relationship occurs among companies of the same group, like in a subsidiary. Finally, *financial interest collaboration* occurs between companies that provide services to other companies. These relationships are illustrated in the figure by the relationships of firms C1, C2 and C3.

In this scenario we can consider different weights for each kind of relationship depending on the analysis to be carried out. For example, if we want to evaluate which companies work together, we could take $\alpha < 0$ for competition relationships, $\alpha > 1$ for partnership collaborative relationships, and $0 < \alpha < 1$ for the collaborative relationships with a financial interest.

Thus, if a new problem has different weights for each relationship ($\alpha \neq 1$), our multi-relational social network model can still be used.

The result of equation (2) was normalized with the application of the natural logarithm and Min-Max normalization, and a connected graph was obtained.

### 3.2 Relationship Age

Another important factor to be considered when defining the relationship degree is age, i.e., we have to know the year that the relationship was created. The relationship age is useful for indicating whether the relationship reflects a connection of present elements, or if it is just a connection that existed in the past.

There are two types of relationships to be considered when looking at the year in which the relationship occurred. The first type is the *exact relationship* that occurs at a given time and the connection between the elements will not necessarily continue over time, like in co-authorship relationships. The *ongoing relationship* is that which

has a defined duration, i.e., it reflects a coexistence of the elements during a time interval, like the relationship during participation in a project.

In order to illustrate the importance of analyzing the relationship age, assume that two researchers A and B published three papers twenty years ago, and two other researchers C and D published one paper last year. If we consider only the number of publications in common, we conclude that A and B have a relationship stronger than researchers C and D. However, relationships between A and B are very old, so that these researchers may not be working together nowadays. Moreover, the relationship between C and D is recent, which leads us to believe that they currently share common interests.

To consider relationship age we added a *year weight* for the relationships in equation (2) and obtained the following equation:

$$TR_{AB} = \sum_{i=1}^{t} \sum_{j=1}^{d} \rho_j \alpha_i R_i, \tag{3}$$

where $d$ is the relationship duration in years and $\rho_j$ is as follows:

$$\rho_j = e^{1/(BY-RY)} \tag{4}$$

$BY$ is the base year used in this work, being the current year (2011), and $RY$ is the year of the relationship. We assume that exact relationships have lasted one year and ongoing relationships have duration equal to the number of years that the relationship has existed.

The function defined in equation (4) describes the curve shown in Figure 4. We can see that the more recent the relationship year is, the greater the weight applied to that relationship. Thus, we guarantee that newer relationships have a bigger weight in the social network analysis than older ones.

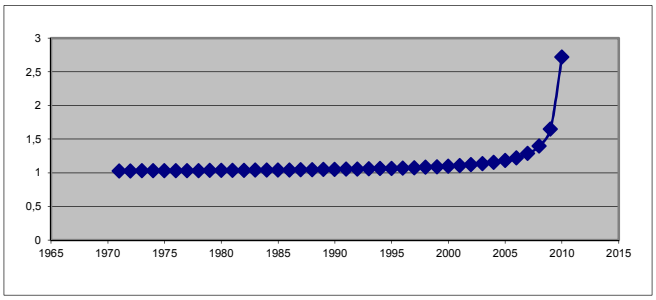


*Figure 4: Graphical representation of equation 4*

After equation (3) we have built an $M$ x $M$ weighted matrix that represents the relationship degree between each pair of researchers, where $M$ is the number of researchers. As the relationship degree represents the similarity between researchers we call this matrix a *similarity matrix*, represented in (5).

$$SM = \begin{cases} TR_{AB} & \text{if researcher } A \text{ links to researcher } B \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

### 3.3 Content Loss in Long Relationships

Another concept introduced in the modeling of our social network is the loss of information when the path between the source and destination node is very large. We believe that all knowledge that is passed from one individual to another has some loss of content.

There are many reasons to explain the loss of knowledge during transfer, such as: mistakes in the information transferred; transferring incomplete information; knowledge misinterpretation; desire to retain part of knowledge acquired for self-protection; dispute for knowledge etc.

Trying to reflect the loss during information exchange, we consider that the receptor receives the information with a loss of *N%* of total knowledge that he could receive, where *N* is the number of intermediary nodes between source and receptor.

The idea is to add a *resistance* to knowledge flow when this knowledge is going through at least one intermediate element, i.e., the path between source and receptor is bigger than one. We aim to bring the multi-relational network model as close as possible to reality. Thus, assuming that the maximum knowledge that can be sent between two researchers A and B is given by $MaxFlow_{AB}$ , then the new relationship degree between them will be given by:

$$\overline{TR} = MaxFlow_{AB} - \frac{N * MaxFlow_{AB}}{100} \quad (6)$$

where *N* is the number of intermediate elements between A and B, and $MaxFlow_{AB}$ is the maximum flow calculated using the *similarity matrix* (equation (5)). At the end of this process we have the *max flow matrix* with resistances, which will be used by the clustering algorithm. The algorithm to calculate the maximum flow will be presented in detail in the next section.

## 4 Cluster Algorithm

Cluster Analyses are techniques that aim to identify data subsets based on the similarity among them [Han, 2006]. Objects of the same subset are more similar among themselves than among objects of different groups.

The clustering methodology is different from the classification techniques. Clustering techniques are unsupervised, i.e., data from the training dataset are not labeled to indicate which class they belong to. Moreover, in some cases, we do not have information on the number of problem classes. In this methodology the similarity between objects is extracted from their structures.

There are basically two clustering techniques: hierarchical and partitional clustering [Jain, 1988]. The hierarchical clustering techniques are useful for small data sets, where you can visually analyze the clustering process step-by-step. In these

algorithms, an element does not change to another group once the group to which the element belongs is defined.

Partitional algorithms can be applied to any data set regardless of the size of the dataset. These methods generate a single partition of the data in an attempt to recover natural groups present in the data. In addition, all groups have a central element, and all elements of a group are more similar to the central element of their own group than to the central elements of other groups.

In hierarchical algorithms, it is not necessary to define the number of groups. On the other hand, the partitional algorithms require the number of groups to be set early in the clustering process. Although defining the number of groups is not trivial, we chose to use a partitional clustering algorithm, since we have no restrictions on the data set size and we can also identify the elements that best represent the groups (central elements).

In this methodology the similarity between objects is extracted from their structures. Generally, cluster methods use a matrix that represents the similarity between objects. In our case, we used a matrix that represents the relationship degree between researchers, i.e., we have used the relationships as a similarity metric. Our similarity matrix was built in the previous section.

### 4.1 Relationships as a Similarity Metric

In previous studies [Stroele, 2009] [Silva, 2009] [Stroele, 2011] we used an algorithm to detect clusters using a minimum spanning tree. This method uses both the profile data and the information on relationships, during the clustering process. However, as this algorithm uses the minimum spanning tree of the graph representing the Social Network, a large part of the relationships is "deleted". Thus, only the stronger relationships are considered in the group detection stage and, consequently, many relationships that can influence the group formation are not considered.

Moreover, it was observed in this algorithm that the relationship is not the criterion with greatest influence in the group detection, as the profile also influences the choice of the group of elements. Thus, researchers who had strong relationships and different profiles, were assigned to different groups. This was the most important factor in the choice of another clustering algorithm for the study of social networks as the scientific relationship is the most important data element for the construction of the groups.

Based on previous research, this work used a clustering algorithm that only evaluates the information contained in relationships to analyze the Scientific Network.

During algorithm development we consider that the Scientific Social Network is a graph called a Social Graph, where each node represents a researcher and each edge represents a relationship between two scientific researchers. The relationship represents an existing social relationship between two people. The social relationship can be strong, between two strong people, or weak between two others. It depends on the measure chosen for the relation.

The goal of the method is to identify groups of people in the social graph that have a strong relationship (good knowledge flow) between them. In order to identify groups of people, the method follows the strategy of analyzing the knowledge flow in social networking. Thus, people who have a large flow of information between them tend to belong to the same group.

The proposed method is based on the problem of maximum flow in networks [Boaventura, 1996]. The weight of each edge of the social graph is a measure of similarity of the nodes connected by those edges. In this method, the weight of edges is set by the attributes of the relationships between nodes. The definition of the weights can be analyzed in more detail in Section 3.

### 4.2 Cluster Algorithm: Max Flow

The Network Flow problem can be interpreted as a Graph Flow problem. The Social Graph with flow will be represented by $\mathbf{G} = (X, U, \mathbf{f})$, in which vector **f** has a dimension *m+1* and can be written thus:

$$\mathbf{f} = (f_0, f_1, \ldots, f_m) \tag{7}$$

Vector **f**, represented in equation (7) is the flow in graph **G** and each component indicates the flow value in a link as part of **G**. The Social Graph is represented in this work by a non-oriented graph. Thus, the max flow going out from $x_i$ to $x_j$ is equal to the max flow from $x_j$ to $x_i$, for $x_i, x_j \in X$.

The clustering algorithm developed in this paper uses the maximum flow between two researchers as a measure of similarity. Whereas social graph **G** represents a Scientific Social Network, in which we have *X* as the set of researchers, *U* as the set of relationships between researchers, and **f** as the set of maximal flows between each pair of researchers. Thus, for all $x_i, x_j \in X$ the maximum flow between these two researchers will be $f_w$, where $0 \le w \le m$.

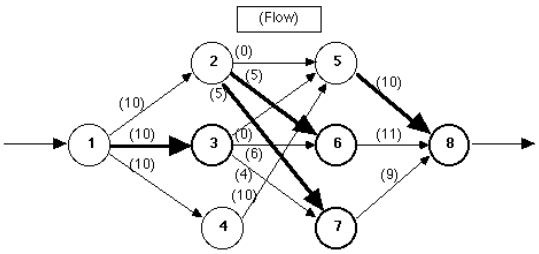


*Figure 5: Maximum Flow Example*

The calculation of the maximum flow between each data set was done using the Edmonds-Karp algorithm [Edmonds, 1972]. This algorithm is a variation of the maximum flow algorithm of Ford-Fulkerson [Ford, 1956]. The main difference between these two implementations is that the Edmonds-Karp algorithm seeks the maximum flow between two elements for the shortest path. Thus, we have the guarantee that the algorithm will converge in a finite number of iterations, even for non-oriented graphs such as the social graph used in this work.

The maximum flow problem is to find the maximum flow possible from some given source node to a given sink node. While we want to know the maximum flow of the graph illustrated in Figure 5, we have to analyze how many units of flow each

node can pass to another. Based on Figure 5, we can see that node 1 can pass up to ten units to nodes 2, 3 and 4. Node 2 can pass up to 5 units to both nodes 6 and 7. Node 3 can pass up to six units of flow to node 6, and four units of flow to node 7. Node 4 can pass up to ten units of flow to node 5. Finally, nodes 5, 6 and 7 can pass up to ten, eleven and nine units of flow to receptor node 8, respectively. Thus, the maximum flow from node 1 to node 8 is 30 units.

In order to join both the max knowledge flow and cluster algorithm, we calculated the maximum flow between all pairs of researchers in the multi-relational scientific social network. The maximum flow computation was done using the matrix built, in which criteria are applied to define weights for relationships (section 3). As we said in the previous section, this matrix represents the similarity degree between researchers. After obtaining the maximum flow between all pairs of authors we built the max flow matrix. This matrix represents the max knowledge flow among researchers and will be used by the clustering algorithm.

The objective of the developed algorithm is to group researchers who have the greatest flow of knowledge between them. We use the k-medoids algorithm [Han, 2006] as a base to help in the development of our algorithm.

In the algorithm developed as well as in the k-medoids algorithm during the first step of the algorithm we randomly define the k medoids and associate each one of them to a group. In the second step, we associate each element of the dataset to the group with which it has the best communication i.e., it is associated with the group with which it has the largest information flow. In the third step, we again define the medoids for each group. The medoids definition in our algorithm is based on the communication skills of each researcher in one's group. We add all the knowledge flows in the group for each researcher, and the researcher who possesses the largest amount is considered the medoids of the group. After defining the new medoids, we move back to step two. As in the k-medoids algorithm, this process continues until there are no changes in the structure of the groups.

In order to assess the optimal number of groups we use the intra-group flow, which is similar to intra-group distance. Cluster validation is an assessment of which cluster has the best clustering structure. This measure is defined as follows:

$$IntraGroupFlow = \sum_{i=0}^{k} \sum_{j=0}^{m} MaxFlow(x_j, \overline{x}_i). \tag{8}$$

where $k$ is the number of groups, $m$ is the number of elements of group $i$, $x_j$ is an element of group $i$, and $\overline{x}_i$ is the medoids of group $i$. Finally, $MaxFlow(x_j, \overline{x}_i)$ is the maximum flow between element $x_j$ and the medoids of your group $\overline{x}_i$.

The goal is to maximize the intragroup flow (Figure 6) in order to obtain the best number of clusters, i.e., the maximum value for this measure indicates the best flow within groups and, consequently, the best partitioning. Based on Figure 6, we chose $k = 31$ as it is the partitioning with the highest intragroup flow.

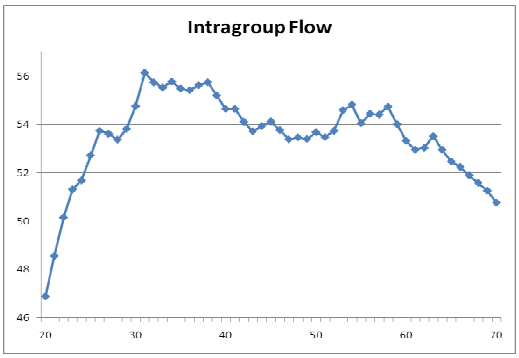


*Figure 6: Intragroup Flow variation*

## 5 Case Study

The case study for this work aims at identifying research communities within and among Brazilian universities using a scientific social network formed by researchers and four different types of relationships that exist in scientific institutions. The data set was described in Section 3. Analyzing the Brazilian scientific social network we aim to learn how information exchange occurs between researchers and their institutions.

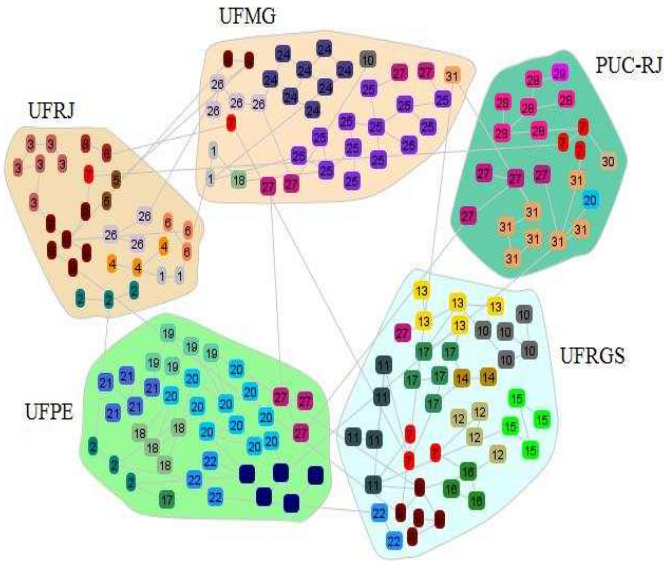


*Figure 7: Inter and Intra institutional relationships*

Figure 7 shows the results obtained by the data mining technique. The largest regions show the Brazilian institutions, and the smaller squares with the same number

show the groups within an organization. Each number in the smaller squares identifies a cluster generated by the group detection method. The edges have information on four relationships as shown in Section 2. In Figure 7 these edges represent only the strongest relationships.

### 5.1 Cluster Analysis

Cluster Analysis allows us to evaluate how communication between researchers occurs. Furthermore, we can study the knowledge exchange amongst groups and educational institutions.

In our first studies we considered only co-authorship relationship [Silva, 2009]. Comparing the results obtained by previous studies and this one we can see that with the addition of new relationships in the social network, the group structure changed. In some cases researchers who were in different groups moved to the same group. This occurred because these researchers have multiple relations with each other. Consequently, the relationships of these researchers are stronger than the relationships of those researchers who have only co-authored relations.

Figure 8 shows the scenario described previously. In Figure 8a, we have an example of the generated groups using only the co-authored relationships. On the other hand, in Figure 8b, we have the groups formed with the inclusion of the new relationships described in Section 2.

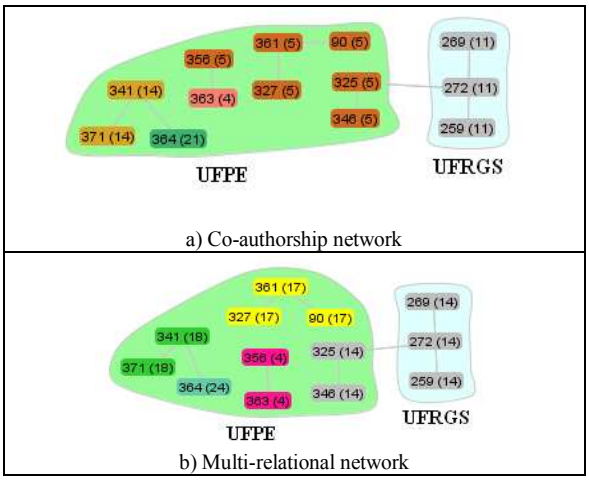


*Figure 8: Cluster Structure Analysis*

Using multi-relational network analysis we can see that researchers 361, 327 and 90 formed a new group, while researchers 325 and 346 moved to the group with researchers 269, 272 and 259. Thus, by building a multi-relational network we came very close to the real world of Brazilian institutions.

As mentioned in Section 4, the analysis of this Scientific Social Network was done in previous work, using another algorithm based on the Minimum Spanning Tree. In analyzing the results obtained with the new algorithm we saw that the

number of Unit Groups was reduced, i.e., the number of groups formed by only one researcher decreased considerably. In previous work [Stroele, 2009] [Silva, 2009] we found ten Unit Groups, whereas we found only two for this work, when using the algorithm that analyzes network flow.

Table 1 was built based on Figure 7 and shows group distribution within and amongst institutions. The diagonal represents the number of groups that exist only in a specific university. Most groups belong to a specific institution; however, some of them have researchers from more than one institution, and in this case we can say that there is a strong exchange of knowledge amongst these institutions.

Looking at Table 1 we can see that the pairs (UFRJ, UFMG) and (UFMG, UFRGS) are the universities with most common groups; see cells (1, 3) and (3, 4), respectively. On the other hand, the pairs (UFRJ, PUC-RJ) and (UFRJ, UFPE) have the fewest number of common groups, i.e. the knowledge flow among these institutions is weaker than in others.

| | **UFRJ** | **PUC-RJ** | **UFMG** | **UFRGS** | **UFPE** |
|---|---|---|---|---|---|
| **UFRJ** | 5 | 1 | 4 | 2 | 1 |
| **PUC-RJ** | 1 | 3 | 3 | 2 | 2 |
| **UFMG** | 4 | 3 | 2 | 4 | 2 |
| **UFRGS** | 2 | 2 | 4 | 6 | 3 |
| **UFPE** | 1 | 2 | 2 | 3 | 3 |
| **COMMON GROUPS** | 8 | 8 | 13 | 11 | 8 |

*Table 1: Group Distribution*

In the analysis of Figure 7 we can see that there are researchers who are responsible for establishing a strong relationship with researchers in external institutions. With the help of these researchers, information can be propagated through the social network more easily, i.e., information from an institution is transferred to another more easily through these researchers. It is important to say that there are many other relationships among institutions than the ones shown in Figure 7. Therefore, the knowledge flow does not only depend on these researchers.

### 5.2 Relationship Analysis

The results allow us to analyze the social network globally and locally. Through a global perspective it is possible to see all the relationships between the educational institutions, by analyzing the knowledge flow among them. For the local analysis we look at specific researchers and analyze how they collaborate with each other.

First we studied the strength of the relationships as established between the educational institutions. We classified them as: internal and external. Relationships between researchers who belong to a single institution are called internal, and external relationships are those that connect two researchers belonging to different institutions.

We examined the number of relationships between each pair of institutions by building a symmetric matrix, shown in Table 2, in which the pair ij represents the total number of relationships between institutions i and j.

| | 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 |
| **TOTAL (RR)** | 16.7 | 28.7 | 14.3 | 17.7 | 16.2 |

*Table 2: Total Number of Relationships*

We examined the total number of external relationships per researcher (RR), because the number of researchers is different for each institution. Therefore, it can be seen that PUC-RJ is the institution that does the most work with other institutions. On the other hand, UFMG is the university that has the fewest researchers linking up with other institutions.

| | UFRJ | PUC-RJ | UFMG | UFRGS | UFPE |
|---|---|---|---|---|---|
| **UFRJ** | – | 1 | 5 | 1 | 1 |
| **PUC-RJ** | 1 | – | 1 | 0 | 2 |
| **UFMG** | 5 | 1 | – | 2 | 1 |
| **UFRGS** | 1 | 0 | 2 | – | 2 |
| **UFPE** | 1 | 2 | 1 | 2 | – |
| **TOTAL** | 8 | 4 | 9 | 5 | 6 |
| **TOTAL (SRR)** | 0.26 | 0.17 | 0.24 | 0.11 | 0.15 |

*Table 3: Number of Strong Relationships*

Table 3 also shows a symmetric matrix where the pair ij represents the total for strong relationships between institutions ‘i’ and ‘j’. Each strong relationship indicates a strong cooperation between researchers.

We examined the total for external strong relationships per researcher (SRR). In examining Table 3, it is possible to see that UFRJ is the institution most strongly linked to other institutions while UFRGS is the least strongly linked.

Comparing the two results presented earlier we can see that there are some researchers who have the profile of working more closely with researchers from other universities. Thus, the relationship between the institutions is strongly linked to a small group of researchers, as can be seen with UFRJ and UFMG. Moreover, 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, as in UFRGS.

As expected, internal relationships are generally stronger than external ones. Thus, it was found that researchers in the same institution have a greater tendency to publish and work together than with those from different institutions.

We also examined the relationships of some specific researchers. The idea was to identify the knowledge brokers. We studied the knowledge brokers of UFRJ. We consider knowledge brokers to be like central connectors: people who, by virtue of their relationships with people in different organizations, serve as boundary spanners (moving information and context from one group to another). Thus, we checked the

amount of inter-group and inter-institutional relationships of each researcher and made a list of them.

From this list we selected three researchers that have different relationship profiles and built a scientific social network with only the relations of researchers 141, 159, and 165. Figure 9 shows a local view of the social network. In this illustration the numbers represent each researcher from the scientific social network.

In Figure 9 we can see that researcher 165 has an internal relationship profile, i.e., this person has more relationships with researchers from the same institution. So we define researcher 165 as an internal central connector.

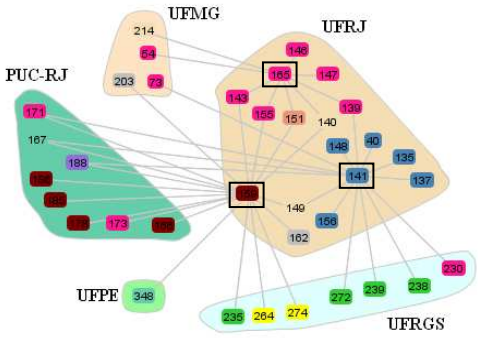


*Figure 9: Local view of the multi-relational social network*

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. It is also possible to see in Figure 9 that researcher 159 has six internal relationships, twelve external ones, and is therefore an external central connector, since he is a sizeable collaborator with other institutions but does comparatively little internal collaboration.

Finally, there is a group of researchers who have internal and external relationships in the same proportion. Researcher 141 who has the same internal and external level of cooperation illustrates this case and can be defined as an internal/external central connector.

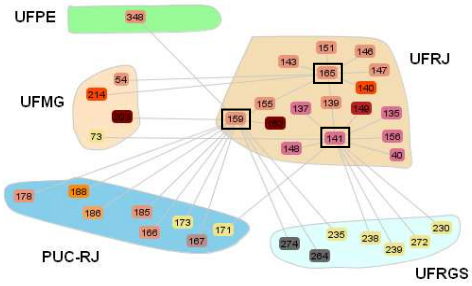


*Figure 10: Local view of Co-authorship Network*

With the local analysis of the social network we could see in detail that there were new knowledge flows in the multi-relational social network. We concluded that sometimes researchers do not have the relationship type represented by the homogeneous social network, but have other kinds of relationships that are represented in the multi-relational social network. This is one great advantage of the heterogeneous social network.

By looking at Figure 10 we can see that there are some researchers who do not have a co-authored relation like subjects 159 and 143 do. However, these researchers have other types of relationships, as shown in Figure 9. Thus, the multi-relational social network has many alternative knowledge flows that are not represented in the homogeneous social network. Consequently, when we compare these two types of networks, we conclude that the multi-relational social network has a better knowledge flow than the homogeneous social networks.

### 5.3 Result Validation

We confirm the validity of the analysis of the relationships and groups formed by the method proposed, using a qualitative evaluation. We interviewed – with the aid of a questionnaire - the researchers from one of the universities, analyzed the answers, and then compared them with the results of our approach.

We requested researchers to indicate how often they write with internal researchers and with external researchers. Thus, it was possible to determine if the researcher had an internal or external relationship profile. This questionnaire was initially distributed only to COPPE/UFRJ researchers.

During this qualitative evaluation, the researchers indicated the areas in which they mostly published, which allowed us to identify interdisciplinary areas. And we found that these researchers are members of the same research group that we identified by using our approach, thus validating the interdisciplinary groups formed by the group detection method.

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 names of the colleagues to whom he is most closely professionally related which allowed strong and weak relationships to be evaluated.

External relationships (COPPE/UFRJ with other universities) were also proven. The results obtained with the questionnaire show that both the methods proposed for analyzing the scientific social network are suitable.

## 6 Conclusion and Future Works

When we improve communication in a Scientific Social Network as a whole, we have more and better interaction between experts from different areas. The improvement can be achieved through the union of experts in a specific area who could work cooperatively. As a result of this interaction, more high-quality data can be generated and, consequently, there is a great improvement in data arranged on the Web.

In this paper we used a group detection method to identify research communities in the Brazilian scientific social network. The result allowed us to make a detailed

analysis of the social network. We looked at various aspects of the social network, such as: level of cooperation between institutions; strong and weak relationships; researchers who play a centralizing role in the social network, etc.

Our next goal is to analyze scientific social network evolution over time. Thus, after understanding how this evolution occurs, we aim to predict and suggest new relationships in order to increase the knowledge flow and improve the information exchange across the scientific social network as a whole.

In future works, we would like to understand more deeply the central connectors and confirm if they are boundary spanners or if some bottleneck researchers exist. Consequently, we would like to know how the analysis of a network reveals the central connectors, the impact of these people on an educational institution, and, finally, the actions that have to be taken to either (1) acknowledge and recognize these people or (2) shift the work patterns to alleviate the bottlenecks.

**Acknowledgements**

This work was supported by CAPES, CNPq, and the COPPETEC Foundation.

## References

[Agrawal, 2003] Agrawal, R., Rajagopalan, S., Srikant, R., Xu, Y. Mining newsgroups using networks arising from social behavior. 12th International WWW Conference. 2003.

[Bird, 2006] Bird, C., Gourley, A., Devanbu, P., Gertz, M., Swaminathan, A. Mining email social networks. Proceedings of International workshop on Mining software repositories. 2006.

[Boaventura, 1996] Boaventura, P.O., Grafos: teoria, modelos, algoritmos. Edgard Blücher LTDA. 1996.

[Cai, 2005] Cai, D., Shao, Z., He, X., Yan, X., Han, J. Community mining from multi-relational networks. in Proceedings of the 2005 European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'05). 2005. Porto, Portugal.

[Cross, 2004] Cross, R.L., A. Parker, and R. Cross, The Hidden Power of Social Networks: Understanding How Work Really Gets Done in Organizations. 2004: Harvard Business Press.

[Edmonds, 1972] Edmonds, J. and R.M. Karp, Theoretical improvements in algorithmic efficiency for network flow problems. Journal of the ACM 19 (2): 248–264, 1972.

[Farrel, 2005] Farrel, S., C. Campbell, and S. Mayagmar. Relescope: An Experiment in Accelerating Relationships. in Conference on Human Factors in Computing Systems. 2005.

[Flake, 2000] Flake, G.W., S. Lawrence, and C.L. Giles. Efficient identification of Web communities. ACM SIGKDD, 2000.

[Ford, 1956] Ford, L.R. and D.R. Fulkerson, Maximal flow through a network. Canadian Journal of Mathematics 8: 399–404, 1956.

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

[GCC, 2006] GCC: A Knowledge Management Environment for Research Centers and Universities, in Lecture Notes in Computer Science. 2006.

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

[Golbeck, 2005] Golbeck, J. Semantic Web Interaction through Trust Network Recommender Systems End User Semantic Web Interaction Workshop. in 4th International Semantic Web Conference. 2005.

[Han, 2006] Han, J. and M. Kamber, Data Mining: Concepts and techniques. 2006.

[Huang, 2005] Huang, Z., X. Li, and H. Chen. Link Prediction Approach to Collaborative Filtering. in Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries. 2005.

[Ichise, 2005] Ichise, R., H. Takeda, and K. Ueyama. Community Mining Tool using Bibliography Data. in Proceedings of the Ninth International Conference on Information Visualisation. 2005: IEEE.

[Jackson, 2010] Jackson, M.O., Social and Economic Networks. 2010: Princeton University.

[Jain, 1988] Jain, A.K. and R.C. Dubes, Algorithms for Clustering Data. 1988, Michigan State University: Prentice Hall.

[Jung, 2007] Jung, J.J., K. Juszczyszyn, and N.T. Nguyen, Centrality Measurement on Semantically Multiplex Social Networks: Divide-and-Conquer Approach. International Journal of Intelligent Information and Database Systems, 1(3/4), 277-292.

[Jung, 2010(a)] Jung, J.J., On Sustainability of Context-Aware Services Among Heterogeneous Smart Spaces. Journal of Universal Computer Science, 16(13), 1745-1760.

[Jung, 2010(b)] Jung, J.J., Integrating Social Networks for Context Fusion in Mobile Service Platforms. Journal of Universal Computer Science, 16(15), 2099-2110.

[Jung, 2010(c)] Jung, J.J., Reusing Ontology Mappings for Query Segmentation and Routing in Semantic Peer-to-Peer Environment. Information Sciences, 180(17), 3248-3257.

[Jung, 2011] Jung, J.J., Service Chain-based Business Alliance Formation in Service-oriented Architecture. Expert Systems with Applications, 38(3), 2206-2211.

[Jung, 2012] Jung, J.J., Evolutionary Approach for Semantic-based Query Sampling in Large-scale Information Sources. Information Sciences, 182(1), 30-39.

[Kumar, 1999] Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A. Trawling the Web for emerging cyber communities. In Proceedings of The 8th International World Wide Web. 1999.

[LATTES, 2006] LATTES, LATTES. 2008, http://lattes.cnpq.br/eng/.

[Liben-Nowell, 2003] Liben-Nowell, D. and J. Kleinberg. The Link Prediction problem for social networks. in Proceedings of the twelfth international conference on Information and knowledge management. 2003. 556-559.

[Lim, 2005] Lim, M., M. Negnevitsky, and J. Hartnett. Artificial Intelligence Applications for Analysis of E-mail Communication Activities. in Proceedings International Conference On Artificial Intelligence In Science And Technology, p.p. 109-113. 2005.

[Newman, 2000] Newman, M.E.J. Who are the best connected scientists? A study of scientific co-authorship networks. 2000. Santa Fe: SFI Working Paper 00-12-64.

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

[Newman, 2001(b)] Newman, M.E.J. Scientific collaboration networks. I. Network construction and fundamental results. 2001: Physical Review E, vol. 64, no. 1, pp. 016 131+.

[Newman, 2001(c)] Newman, M.E.J. Scientific collaboration networks. II. shortest paths, weighted networks, and centrality. 2001: Physical Review E, vol. 64, no. 1, pp. 016 132+.

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

[Newman, 2004(b)] Newman, M.E.J. Co-authorship networks and patterns of scientific collaboration. in Proceedings of the National Academy of Sciences, 101: 5200-5205. 2004.

[Pioch, 2005] Pioch, N., Barlos, F., Fournelle, C., Stephenson, T. A Link and Group Analysis Toolkit (LGAT) for Intelligence Analysis. 2005: https://analysis.mitre.org/proceedings/Final_Papers_Files/348_Camera_Ready_Paper.pdf.

[Raab, 2003] Raab, J. and H. Milward, Dark Networks as Problems. Journal of Public Administration Research and Theory, vol. 13, no. 4. pp 413-439, 2003.

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

[Silva, 2009] Silva, R., Stroele, V., Oliveira, J., Souza, J. M., Zimbrao, G. Mining and Analyzing Organizational Social Networks for Collaborative Design. in 13th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2009). 2009.

[Stroele, 2009] Stroele, V., Oliveira, J., Zimbrao, G., Souza, J. M. Mining and Analyzing Multirelational Social Networks. in 2009 International Conference on Social Computing. 2009. Vancouver: Proceedings of International Conference on Social Computing (IEEE CS).

[Stroele, 2011] Stroele, V., Silva, R., Souza, M. F., Mello, C. E. R., Souza, J. M., Zimbrao, G., Oliveira, J. Identifying Workgroups in Brazilian Scientific Social Networks. Journal of Universal Computer Science, 2011.Vol. 17, No. 14: p. 1951-1970.

[Tyler, 2003] Tyler, J.R., D.M. Wilkinson, and B.A. Huberman. Email as Spectroscopy: Automated Discovery of community Structure Within Organizations. in Proceedings of the First International Conference on Communities and Technologies. 2003.

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

[Zhu, 2003] Zhu, J., Mining Web Site Link Structures for Adaptive Web Site Navigation and Search. 2003, University of Ulster: http://kmi.open.ac.uk/people/jianhan/thesis.pdf.