Potential Collaboration Discovery using Document Clustering and Community Structure Detection

Cristian K. dos Santos
COPPE/UFRJ
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
+552125627388
c.klen@coc.ufrj.br

Alexandre G. Evsukoff
COPPE/UFRJ
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
+552125627388
evsukoff@coc.ufrj.br

Beatriz S.L.P. de Lima
COPPE/UFRJ
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
+552125627388
bia@coc.ufrj.br

Nelson F. F. Ebecken
COPPE/UFRJ
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
+552125627389
nelson@ntt.ufrj.br

ABSTRACT
Complex network analysis is a growing research area in a wide variety of domains and has recently become closely associated with data, text and web mining. One of the most active areas in the study of complex networks is the detection of community structure, which can be related to the clustering problem in data mining. This paper employs a community structure detection algorithm for document clustering in order to discover potential relationships in a social network. The proposed approach is explored in a case study of potential collaboration discovery among the research staff in the Graduate Civil Engineering Department of the Federal University of Rio de Janeiro, Brazil. The results show that the combined use of both techniques provides useful insights on the relationships, both existent and potential, among individuals in the social network.

Categories and Subject Descriptors
H.3.4 [Systems and Software]: Information networks, H.2.8 [Database Applications]: Data Mining; H.3.3 [Information Search and Retrieval]: Clustering; I.7.5 [Document Capture]: Document analysis.

General Terms
Algorithms, Management, Documentation, Human Factors.

Keywords
Community structure detection, complex networks, text mining, documents clustering, spectral clustering.

1. INTRODUCTION
The Internet has allowed social networking to become a worldwide phenomenon that integrates people who would probably never be connected through their conventional social acquaintances.

Individuals are very willing to express themselves in online social networks. The ways they understand the world are expressed by their opinions and thoughts, which are shared, with more or less agreement, with many others. Most content is released in the form of text in a wide variety of formats, which can be useful for the discovery of potential links in a social network. The discovery of potential relationships in a social network stimulates the integration of people and also enhances the information flow in the network.

One of the most active areas in the study of complex networks is the detection of community structure [1]-[8]. The graph theoretic approach is widely used for community structure detection in complex networks and is also the base formalism for spectral clustering [9]-[11], so it is a natural way to integrate these techniques.

The concept of a good clustering or community partition is very difficult and can be formally defined in many ways, so that many different algorithms can be derived. In the graph theoretic approach, the algorithms are usually formulated as graph partition problems, in which the weight of each edge is the similarity between points that correspond to vertices connected by the edge. The goal of this algorithm is to find the minimum weight cuts in the graph, which is a combinatorial problem. The problem is thus usually addressed through spectral decomposition techniques, as described in recent excellent reviews on the subject [10][11].

The most popular class of methods to detect communities is, perhaps, the maximization of the function known as "modularity," introduced by Newman and Girvan [1]. This measure is by far the most used and best-known function to quantify the "goodness" of possible subdivisions of a given network into communities [1]. The modularity measure is, however, not able to detect very small communities, as it has been recently pointed out [6].

Community detection algorithms have been recently studied for document clustering [12][13], where the Newman algorithm [3] was compared to spectral clustering techniques. The Newman algorithm produced better results.

The main contribution of this work is methodology to integrate document clustering and community detection for the discovery of potential relationships in a social network. The results are explored in a case study of potential collaboration discovery in the Graduate Civil Engineering Department of the Federal University of Rio de Janeiro. The Newman algorithm is used both for community detection in the co-authorship network and for document clustering. The co-authorship network is obtained from the collaboration of Professors in MSc and PhD thesis supervision. The corpus used to identify potential links in this

CNIKM'09, November 6, 2009, Hong Kong, China.

network is the set of abstracts of MSc and PhD theses produced in the department from 2004 to 2008.

The problem of potential link discovery in networks has being studied in the recent literature [14]. In the Relational Topic Model [15], it is also possible to predict links using texts' content. The authors have also developed a method to uncover the relationships encoded in a collection of texts using an approach based on a probabilistic topic model [16], which allows inferring descriptions of the network's elements. Kemp et al. [17] have presented an approach to cluster one or more sets of entities and discover the relationships between clusters that are possible or likely.

In this work, as so as in related approaches [14-17], use not only the link structure, but also the features of the network's elements, in this case textual documents, in order to analyze the network for link prediction or discovery. This seems to be more effective as additional information is included into the analysis instead os using only the link structure to predict links [18].

The paper is organized as follows. The proposed methodology is presented in the next section. The modeling of a document collection as a graph and the formulation of the document clustering as a graph cut problem are presented in section three. In section four the Newman algorithm is introduced, and in section five the case study is discussed. The paper finishes with conclusions and future studies in section six.

2. POTENTIAL LINK DISCOVERY

As is typical in complex network analysis, a co-authoring network definition starts with a bi-partite graph defined over two sets of objects [23]. In this work there are three sets of objects, as shown in Figure 1. The first set is the set of individuals of the social network under study. Each document in the set of documents is related to one or more individuals. The set of terms may be a set of keywords within a controlled dictionary or, more generally, the set of generic terms appearing in the documents.

In the case study presented in section 5, the set of individuals are the permanent staff and external collaborators of the Graduate Civil Engineering Department of the Federal University of Rio de Janeiro. The set of documents are the abstracts of the MSc and PhD theses produced in the department from 2004 to 2008. The relationships between individuals and documents are the collaborations in thesis supervision. Each document can have up to three supervisors. The set of terms are the words (stems) appearing in the documents.

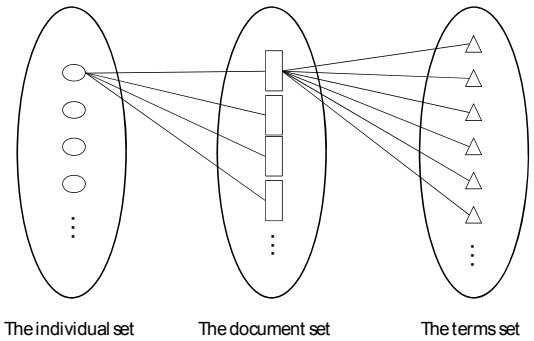


Figure 1. The sets of objects for the definition of the networks in the proposed method

Three kinds of networks can be defined from the object sets shown in Figure 1. The first one is the collaboration network, which is a kind of co-authorship network, in which the nodes are the individuals and a link represents the co-supervision between two individuals in at least one document. This is the base social network used in this study and represents the existing relationships among the individuals in their social interactions. In a more general setting, the existing relationships can also be represented by other kind of networks, such as hyperlinks in a set of blogs, friendship in a social network web service, or emails. The base network may also be weighted or unweighted, directed or undirected. In the case study presented in this work, the base network is un-weighted and undirected.

The second network is the document network, in which the nodes represent the documents and the links represent the similarities among documents, as determined by the terms appearing in the documents. This network is generated artificially using a threshold in the similarity value representing a strong similarity between two vertices. The document network is weighted and undirected.

The third network is a combination of the first two, in which the nodes are the individuals and each link represents the similarity between the content produced by the two individuals. The community structure in this network reveals groups of individuals interested in the same subjects. The comparison of the network structure found for this network with the structure of the base network makes it possible to reveal potential relationships that are not yet present in the base network.

The definition of the document network plays a central role in the methodology described in this work and is discussed in the next section.

3. SPECTRAL CLUSTERING IN DOCUMENT NETWORKS

Unstructured information in document databases presents intrinsic characteristics such that data mining algorithms must be adapted to solve text-mining tasks. The most usual representations for text mining rely on the vector space model of documents, usually in information retrieval [19]. In such a model, the order of words is not considered, and each document in a collection is represented by a vector, in which the components are related to relevant words appearing in the document collection.

In the vector space model, the document collection is represented as the $n \times m$ sparse matrix $\mathbf{X}$, in which the lines are related to the documents and the columns are related to the terms. An element x_{ij} accounts for how the term T_i is related to the document D_j, often computed by the tf-idf frequency [20].

3.1 The document network

A document collection can be viewed as a complex network, in which the nodes are the documents and the edges are weighted according to document similarities.

The document network can be defined from $\mathbf{X}$ as a weighted and undirected proximity graph $G(V,E)$, in which the set of vertices $V = \{v_1,...,v_n\}$ corresponds to the n documents and the set of edges E is defined through the symmetric adjacency matrix $\mathbf{A} \in \Re^{n \times n}$. Each element entry $a_{ij} \in \mathbf{A}$ represents the pair-wise similarity between the documents D_i and D_j, computed as:

$$a_{ij} = \begin{cases} h(\mathbf{x}_i, \mathbf{x}_j) & \text{if } i \neq j \text{ and } h(\mathbf{x}_i, \mathbf{x}_j) \geq \varepsilon, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The function h measures the local neighborhood relationships between two vertices, which should be greater than or equal to the parameter ε that defines the radius of proximity among the documents. This parameter is very important in the definition of the document network structure since $\varepsilon = 0$ defines a complete network. Different results are obtained with different values of ε; this issue is further exploited in section 5.

The similarity function h can be computed by different functions. In spectral and kernel clustering literature, the Gaussian similarity function is usually employed. In text mining applications, the cosine similarity function is usually employed within the vector space model [20]. The cosine similarity function has shown good results in previous studies of document clustering within a framework of spectral clustering [22]. It is defined as:

$$h(\mathbf{x}_i, \mathbf{x}_j) = \frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{\sqrt{\langle \mathbf{x}_i, \mathbf{x}_i \rangle \langle \mathbf{x}_j, \mathbf{x}_j \rangle}} \quad (2)$$

where $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j$ is the scalar product.

3.2 Spectral clustering

The graph cut problem aims to separate a subset of vertices $S \subset V$ from its complement $V - S$ denoted by $\overline{S}$ [9][21]. The graph cut problem can be formulated in several different ways, depending on the choice of the objective function to be optimized [11][21]. One of the options is the cut function, whose minimization favors partitions containing isolated vertices. It is defined as follows:

$$cut(S, \overline{S}) = \sum_{i \in S, j \in \overline{S}} a_{ij} \quad (3)$$

To overcome the weakness of the cut function and achieve better balance between partitions, it is recommended to use its normalized version, that is, the normalized cut function [21]:

$$Ncut(S, \overline{S}) = cut(S, \overline{S}) \left(\frac{1}{vol(S)} + \frac{1}{vol(\overline{S})} \right) \quad (4)$$

where $vol(S)$ is the volume of S, computed as:

$$vol(S) = \sum_{i \in S} d_i \quad (5)$$

The degree d_i of a vertex $v_i \in V$ is the number of edges incident to the vertex and is defined as:

$$d_i = \sum_{j=1}^{n} a_{ij} \quad (6)$$

The minimization of the function (5) is an NP-hard problem that can be relaxed by introducing the graph Laplacian matrix [11][21].

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \quad (7)$$

where the degree matrix $\mathbf{D}$ is defined as the diagonal matrix of the degrees $d_1, ..., d_n$.

The graph Laplacian is a positive semi-definite matrix, such that its eigenvalues are always positive real-valued. Some spectral clustering approaches are based on the solution of the generalized eigenvalue problem [1]:

$$\mathbf{LU} = \mathbf{\Lambda DU} \quad (8)$$

where $\mathbf{\Lambda}$ is the diagonal matrix of the eigenvalues, which are ordered in ascending order, $0 \leq \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$. The orthogonal matrix $\mathbf{U}$ is the matrix in which the columns are the generalized eigenvectors.

Good clustering algorithms for graph clustering depend on the quality of the objective function being used. Recently, a cost function called the modularity function was proposed by Newman and Girvan, [1] to overcome limitations of the previous measures for measuring community structure, as discussed in the next section.

4. MODULARITY-BASED COMMUNITY DETECTION

4.1 The modularity and the community structure in networks

A community structure in a network $G(V, E)$ is defined as a partition P_K of the set of vertices into K subsets $C_j, j = 1...K$, such that $\bigcap_{j=1...K} C_j = \varnothing$ and $\bigcup_{j=1...K} C_j = V$.

One can think about group structure in graph clustering problems as clusters with high density of edges within them, and a lower density of edges among them.

Newman and Girvan [1] defined a quantitative measure called modularity to evaluate an assignment of nodes into communities. This measure can be used to compare different assignments of nodes into communities. The network modularity Q is defined over a network partition P_K as:

$$Q(P_k) = \sum_{j=1}^{K} \left(\frac{R(C_j, C_j)}{R(V, V)} - \left(\frac{R(C_j, V)}{R(V, V)} \right)^2 \right) \quad (9)$$

where $R(C', C'') = \sum_{i \in C', j \in C''} a_{ij}$ measures the association among the nodes of the subsets C' and C''. Thus, $R(C_j, C_j)$ measures the within-community sum of edge weights; $R(C_j, V)$ measures the sum of weights over all edges attached to nodes in community C_j; and $R(V, V)$ is the normalization term that measures the sum over all edge weights in the entire network.. Considering binary weights, the first term $R(C_j, C_j) / R(V, V)$ is the empirical probability that both vertices of a randomly selected edge fall in subset C_j. The second term $\left(R(C_j, V) / R(V, V) \right)^2$ is the empirical probability that only one of the ends (either one) of a randomly selected edge falls in subset C_j. Thus, the modularity measures the deviation between observed cluster structure and what could be expected under an independent random model. If the number of within-community edges is no better than random, then the value $Q = 0$. A value of $Q = 1$, which is the maximum,

indicates strong community structure. In practice, however, values typically fall in the range from 0.3 to 0.7 [1].

In the next subsection, the modularity function will be reformulated as a spectral optimization problem, according to Newman [3].

4.2 The spectral modularity optimization method

Consider the graph $G(V,E)$, and suppose a particular partition of G into two groups $S \subset V$ and its complement $V-S$, denoted by $\overline{S}$. The partition is defined by the partition vector $\mathbf{q} = (q_1,\ldots,q_n)$, such that $q_i = 1$ if vertex $v_i \in S$ and $q_i = -1$ if vertex $v_i \in \overline{S}$.

The expected edge weight p_{ij} between vertices v_i and v_j when edges are placed at random is computed by [3]:

$$p_{ij} = \frac{d_i d_j}{2m} \tag{10}$$

where d_i and d_j are the degrees of the vertices v_i and v_j as defined by (3), and m measures the sum of all edge weights in the entire network:

$$m = \frac{1}{2}\sum_{i=1..n} d_i \tag{11}$$

The modularity measure Q can be written as the sum of the differences between a_{ij} and p_{ij} over all pairs of vertices v_i and v_j that fall in the same community:

$$Q = \frac{1}{4m}\sum_{\substack{i=1\ldots n \\ j=1\ldots n}} (a_{ij} - p_{ij}) q_i q_j \tag{12}$$

which is written in the matrix format as:

$$Q = \frac{1}{4m}\mathbf{q}^{\mathrm{T}}\mathbf{B}\mathbf{q} \tag{13}$$

where $\mathbf{B}$ is a real and symmetric matrix, called the modularity matrix, in which the elements are computed as:

$$b_{ij} = a_{ij} - p_{ij} \tag{14}$$

The maximization of the modularity (13) is equivalent to a graph cut problem such that an approximate solution can be computed by the spectral decomposition of $\mathbf{B}$:

$$\mathbf{B}\mathbf{z} = \mathbf{z}\beta \tag{15}$$

where $\mathbf{z}$ is the eigenvector corresponding to the largest eigenvalue β. The approximate solution corresponds thus to the maximization of the Rayleigh quotient:

$$\hat{Q} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{B}\mathbf{z}}{\mathbf{z}^{\mathrm{T}}\mathbf{z}} \tag{16}$$

A partition of the network is computed by maximizing the modularity $\hat{Q}$ by choosing appropriate values for the partition vector $\mathbf{q}$ according to the sign of the components of the eigenvector $\mathbf{z}$:

$$q_i = \begin{cases} +1 & if\ z_i \geq 0 \\ -1 & if\ z_i < 0 \end{cases} \tag{17}$$

The partition vector defined by (17) divides the network into only two communities. However, many networks contain more than two communities. In such cases, this approach can be applied recursively to find a partition of the network into more than two communities.

The idea of the recursive algorithm is to evaluate the gain in the modularity function if a community is further divided. For each group C' generated by a partition like (17), the additional contribution to the modularity ΔQ is computed as:

$$\Delta Q = \frac{1}{4m}\mathbf{q}'^{\mathrm{T}}\mathbf{B}'\mathbf{q}' \tag{18}$$

where $\mathbf{B}'$ is the matrix corresponding to the vertices that belong to the group C', and $\mathbf{q}'$ is the partition vector that will subdivide the group C'.

The group modularity matrix $\mathbf{B}'$ is computed as the sub-matrix of $\mathbf{B}$ corresponding to the vertices that belong to the group C', as:

$$b'_{ij} = b_{ij} - \delta_{ij}\sum_{k\in C'} b_{ik} \tag{19}$$

where $b_{ij} \in \mathbf{B}$ are the elements of the modularity matrix computed as in (14), and $b'_{ij} \in \mathbf{B}'$ are the elements of the modularity matrix corresponding to the partition of the group C', where the (i,j) indexes refer to the nodes of the entire network, and δ_{ij} stands for the Kronecker δ.

The recursive process is halted if there is no further division of a subnetwork that will increase the modularity of the network, and therefore there is no gain in continuing to divide the network. In practice, the test $\Delta Q > 10^{-3}$ is used as the stopping criterion.

5. RESULTS

The case study was conducted based on the recent research production of the Graduate Civil Engineering Department of the Federal University of Rio de Janeiro. A set of 147 researchers among members of the staff and external collaborators were involved in the supervision of 585 MSc and PhD theses from 2004 to 2008. The theses' abstracts were collected from the department's website (http://www.coc.ufrj.br/en) and used as the set of documents in the present case study. This case was chosen for study because the authors are members of the staff and interact directly with most of the individuals in the network, allowing the results of the methodology to be easily verified.

The department was formerly organized into the three conventional research areas of Civil Engineering: Structures and Materials (SM), Geotechnical Engineering (GE), and Water Resources (WR). In 2006, as a result of strong interdisciplinary research, four new research areas were added: Oil and Gas (OG), Computational Mechanics (CM), Computational Systems (CS), and Environmental Engineering (EE). A description of the research areas can be found in the department website.

The members of the permanent staff can work in any of the research areas in the department, but each is usually more active

in one main research area. The external researchers may participate in supervision only in collaboration with a member of the permanent staff. Their research areas were considered as the main research areas of their partners.

The next subsections present the results of the community structure detection computed by the Newman algorithm in the three networks generated for the study. The results have been analyzed qualitatively since the numerical evaluation of the Newton algorithm have be presented in previous works [12][13].

5.1 Co-supervision Community Structure

The co-supervision network is defined as a co-authorship network. The nodes are the permanent staff members and the external collaborators, and a link between two people exists if they have worked together on at least one thesis as co-supervisors during the period used in the study (from 2004 to 2008). The co-supervision network thus represents the actual social network of the collaborations in the department.

Community structure detection in the co-supervision network was computed by the Newman algorithm, as described above, and 10 communities were found, as shown in Figure 2(a). The classification of the network nodes according to the main research area of the individuals is shown in Figure 2(b). The network layout was computed by the Spring Layout algorithm, and the colors were kept the same when possible.

The network presents a giant component containing most of the individuals, and the communities identified correspond roughly to the main research areas of the individuals. A closer look at the community structure reveals that the algorithm was able to find most of the research groups in the department.

5.2 Document Clustering

Document clustering is obtained by looking for communities in the document network. The process of finding communities in the document network is an analysis task and aims to find some underlying structure in the document collection.

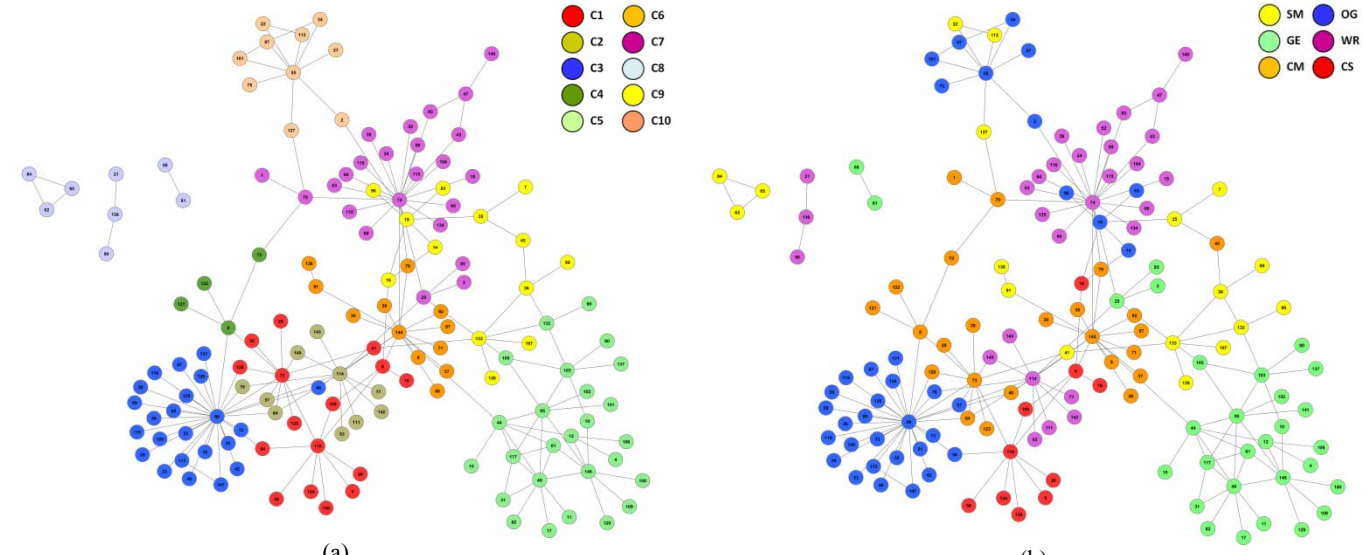
The network structure is obtained by applying a threshold to the similarity function (1). The best structure is searched, varying the parameter ε in (1). In this work, ten runs were computed, varying ε from 0.10 to 0.55. The number of communities (document clusters) obtained and the final value of the modularity parameter (Q) are presented in Table 1.

The analysis of the results shows that the threshold value $\varepsilon = 0.20$ results in seven communities that are very close to the actual research areas in the department.

Table 1. The results of the analysis runs

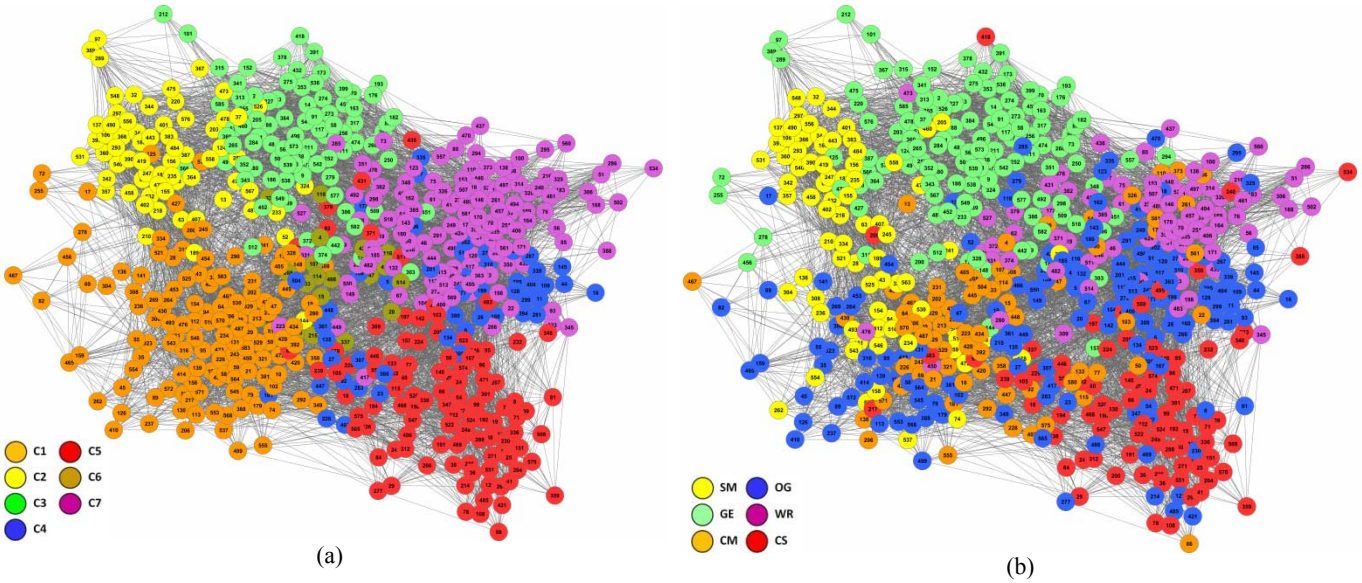
ε	# groups.	Q
0.10	4	0.3321
0.15	5	0.4336
0.20	**7**	**0.5334**
0.25	8	0.6197
0.30	10	0.6978
0.35	13	0.7785
0.40	14	0.8398
0.45	20	0.8882
0.50	28	0.9178
0.55	38	0.9269

The documents can be classified according to the supervisors' main research areas. The visualization of the document network, showing the communities computed by the algorithm and the actual classification, is shown in Figure 3. The colors and positions of nodes were kept constant to help in the interpretation of the result. The visualization was produced by the Cytoscape open source software using the Spring Layout algorithm.



(a) (b)

Figure 2. Community structures in the collaboration network (a) results obtained by the Newman algorithm (b) clustering defined by the main research area of each individual.



(a) (b)

Figure 3. Community structures in the document network (a) results obtained by Newman algorithm (b) classification defined according to the main research area of the permanent staff supervisor.

The analysis of the main terms found in each cluster reveals interesting insights on the recent work produced in the department. Table 2 presents the description of the seven groups found by the algorithm as having the most related research areas, which were interpreted by the analysis of the most relevant terms in each group.

The three classical research areas of the department (Structures and Materials (SM), Geotechnical Engineering (GE), and Water Resources (WR)) have been shown to produce research works very related each to other that could be fairly well identified as clusters. In the case of the recent interdisciplinary research areas, the Computational Systems (CS) area has resulted in a good cluster. The other three areas, Oil and Gas (OG), Computational Mechanics (CM), and Environment Engineering (EE), were more difficult to identify, as they deal with highly interdisciplinary research.

Two of the main research lines in the OG research are Geochemistry and Geophysics simulations, which have very different terms and thus resulted in two different communities. The EE research area could not be identified since, in most of the cases, it is an application of techniques that are developed by other areas. On the other hand, the CM area produces tools that are widely used in the other areas, but it can still be roughly recognized as a group.

Table 2. The interpretation of groups found

Group	Research Area
C1	Computational Mechanics
C2	Structures and Materials
C3	Geotechnics / Environment
C4	Oil and Gas (Geochemistry)
C5	Computational Systems
C6	Oil and Gas (Geophysics)
C7	Water Resources / Environment

For evaluation of the clustering performance, the documents were each classified into one of the research areas of the department according to the main research area of the permanent staff supervisor. The EE researchers were distributed between GE and WR, according to their research areas in the former Department structure. The groups C4 and C6 were aggregated, as they are both related to the OG research area.

The resulting confusion matrix is shown in Table 3, where the actual classification is presented in the lines and the classification proposed by the algorithm is shown in the columns.

Table 3. Confusion matrix

	C1	C2	C3	C4 + C6	C5	C7
CM	**37**	1	0	15	7	13
SM	32	**40**	2	2	1	1
GE	12	22	**80**	1	1	9
OG	42	1	1	**47**	16	39
CS	3	0	1	1	**72**	7
WR	3	1	2	3	9	**61**

It can be seen that documents classified as group C1 have been produced by researchers of four different research areas. The documents classified as group C2 have good accuracy (approximately 62%) with respect to the SM area and are highly confused with documents produced by the GE researchers. The group C3 is the one that best matches a single research area, resulting in 93% accuracy with respect to the theses supervised by GE staff. The groups C4 and C6 together have a good match with the OG research area (68%) and also have a strong connection with the theses produced by the CM researchers. The group C5 presents a good degree of matching (68%) with the CS staff and also a strong connection with the OG researchers. Finally, group C7 matches the research of WR, with some connections to OG and CM.

The remarks and conclusions stated above correspond well to the knowledge of the research work being developed in the department. The proposed approach, to use several slices of network structure as defined by the parameter ε in (1) with the Newman algorithm for document clustering, has shown good results as an analysis tool. The interpretation of the results would be difficult to achieve without the domain knowledge, but that is the case in many clustering applications.

5.3 Potential Collaboration

The potential collaboration network is a synthesis of the previous two networks. The potential collaboration network is a kind of document network where the nodes are individuals, and a link between two individuals represents the similarity of their research subjects.

The first step in the definition of the collaboration network is to compute the vector space term representation matrix $\mathbf{Y}$, in which the lines represent the researchers and the columns represent the terms found in the documents. An element y_{ij} accounts for how the term is related to the researcher, computed by the sum of the term frequencies of all documents that were supervised by that researcher:

$$\mathbf{y}_i = \sum_{j \in \Omega_i} \mathbf{x}_j \qquad (20)$$

where Ω_i is the set of all documents produced by the researcher i.

The community structure detection in the collaboration network is performed in the same way as in the document network. The search for structure was performed by varying ε from 0.10 to 0.55. The number of communities (document clusters) obtained and the final value of the modularity parameter (Q) are presented in Table 4.

Table 4. The results of the analysis runs

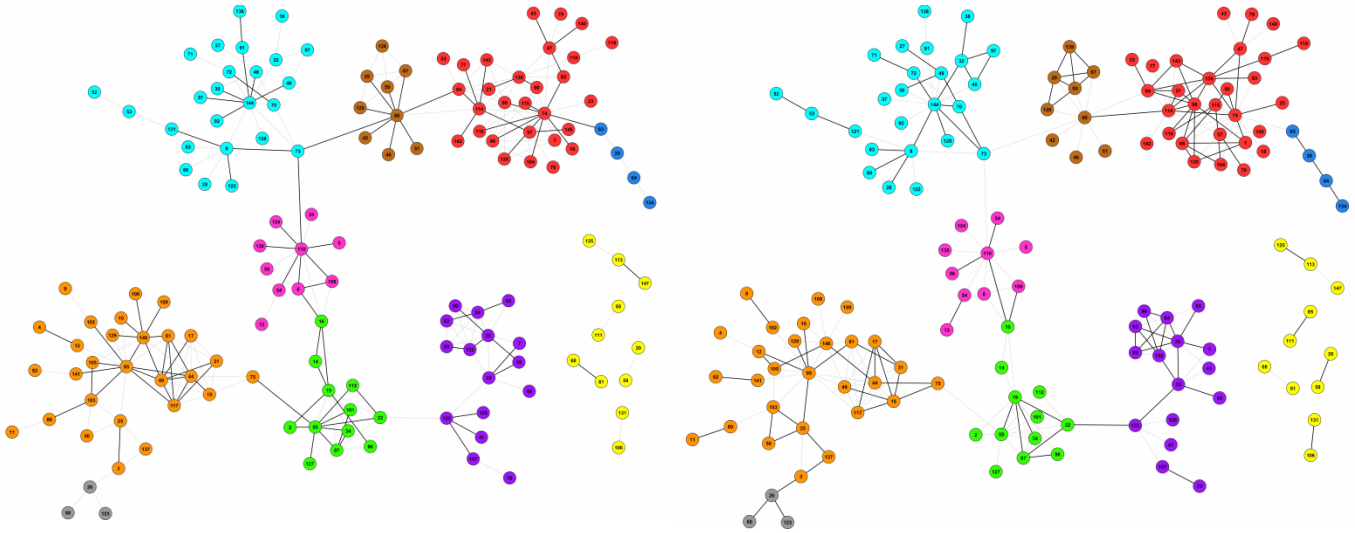
ε	# groups	Q
0.10	4	0.1962
0.15	4	0.2619
0.20	4	0.3476
0.25	5	0.4161
0.30	7	0.5452
0.35	8	0.6546
0.40	8	0.7168
0.45	6	0.7252
0.50	**10**	**0.7913**
0.55	9	0.8304

The community partition corresponding to $\varepsilon = 0.50$ was the one that found the same number of communities as the co-supervision network. The resulting network and its community structure is shown in Figure 4.

Two researchers that have been classified in the same community in the collaboration network may be co-workers or not, depending on whether they have a link in the co-supervision network.

Figure 4(a) highlights the existing links in the collaboration network, i.e., the links that are also present in the co-supervision network. Figure 4(b) highlights the links in the collaboration network that are not present in the co-supervision network and thus represent potential co-workers.

It is possible to suggest the most likely research subject on which the potential co-workers could collaborate by analyzing the most important terms in the collaboration network link.



(a) (b)

Figure 4. Community structures in network of potential collaboration (a) existing links (b) non-existing (potential) links.

The communities in the potential collaboration network represent researchers interested in the same subject. Analysis of the most relevant term in each community allows the identification of the main research area in each community, as shown in Table 5. Table 5 also presents the number of individuals in each community.

Table 5. The interpretation of groups found

Group	#	Research Area
C1	28	Geotechnical Engineering
C2	3	Water Resources/ Environment
C3	12	Oil and Gas / Structures
C4	16	Structures and Materials
C5	25	Computational Mechanics
C6	11	Geotechnical Engineering
C7	10	Computational Systems
C8	9	Oil and Gas
C9	4	Water Resources/ Environment
C10	29	Water Resources/ Environment

6. CONCLUSIONS

This work has presented an approach to identifying potential collaborations in a social network, the individuals of which produce documents. The method is based on document clustering using a network community detection algorithm, which allows definition of the number of clusters, a recurrent problem in cluster analysis. Moreover, the method is recursive so that it can deal with hierarchical structure, a feature frequently found in document clustering problems.

The method was applied in a real case study of the social network of the research staff in the Civil Engineering Department of the Federal University of Rio de Janeiro, Brazil. This network is a complex, highly adaptive dynamic system that cooperates on several research themes. The results obtained with the proposed method have allowed to a better understanding of the research work being produced and also have revealed some unexpected relationships.

This research project will continue in the direction of scalability of the present analysis so that it can be applied in a wider network.

7. ACKNOWLEDGMENTS

The financial support for this research has been provided by the Brazilian Research Agencies, CNPq, CAPES and FINEP. The authors are also grateful to the Graduate Civil Engineering department of the Federal University of Rio de Janeiro.

8. REFERENCES

[1] Newman, M. E. J., and M. Girvan, "Finding and evaluating community structure in networks," Phys. Rev. E 69, 026113, 2004.

[2] Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), v.74, n.3, p.036104-19. 2006.

[3] Newman, M. E. J. Modularity and community structure in networks. PNAS, v.103, n.23, p.8577-8582. 2006.

[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.U. Hwang, "Complex networks: Structure and dynamics," Physics Reports, pp. 175-308, 2006.

[5] E.A. Leicht, and M.E.J. Newman, "Community structure in directed networks," arXiv:0709.4500v1 2007.

[6] Fortunato, S. e M. Barthélemy. Resolution limit in community detection. PNAS, v.104, n.1, p.36-41. 2007.

[7] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," Nature 435, pp. 814-818, 2005.

[8] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Defining and identifying communities in networks, Proc. Natl. Acad. Sci. USA 101, 2658-2663, arXiv:cond-mat/0309488v2, 2004.

[9] Schaeffer, S. E. Graph clustering. Computer Science Review, v.1, n.1, p.27-64. 2007.

[10] Filippone, M., F. Camastra, et al. A survey of kernel and spectral methods for clustering. Pattern Recognition, v.41, n.1, p.176-190. 2008.

[11] von Luxburg, U. A tutorial on spectral clustering. Statistics and Computing, v.17, n.4, p.395-416. 2007.

[12] dos Santos, C. K., Evsukoff, A. G., de Lima, B. S .L. P. Cluster analysis in document networks. WIT Transactions on Information and Communication Technologies, v.40, p.95-104. 2008.

[13] dos Santos, C. K., Evsukoff, A. G., de Lima, B. S. L. P. Spectral clustering and community detection in document networks. WIT Transactions on Information and Communication Technologies, v.42, p.41-50. 2009.

[14] 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, New Orleans, LA, USA, 2003.

[15] Chang, J. and D. M. Blei: Relational topic models for document networks. Proc. of Conf. on AI and Statistics (AISTATS'09), 2009.

[16] Chang, J; J. Boyd-Graber and D. M. Blei: Connections between the lines: augmenting social networks with text. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, (KDD'09) Paris, France, 2009.

[17] Kemp, C.; J. B. Tenenbaum, T. L. Griffiths, T. Yamada and N. Ueda: Learning systems of concepts with an infinite relational model. In AAAI, 2006.

[18] Xu, Z.; V. Tresp, K.Yu and H.-P. Kriegel: Infinite Hidden Relational Models, Proc. 22nd Conf. on Uncertainty in Artificial Intelligence (UAI'06), Cambridge, MA, 2006.

[19] W.B. Michael, D. Zlatko, and R.J. Elizabeth, "Matrices, Vector Spaces, and Information Retrieval," SIAM Rev, pp. 335-362. 1999.

[20] M. Berry, Survey of Text Mining: Clustering, Classification, and Retrieval, Springer, 2003.

[21] F.R.K. Chung, Spectral Graph Theory, CBMS Regional Conf. Series in Mathematics., no. 92. American Mathematic Society, 1997.

[22] Bao, L., et al., Document Clustering Based on Spectral Clustering and Non-negative Matrix Factorization, in New Frontiers in Applied Artificial Intelligence. Springer. p. 149-158, 2008

[23] Jean-Loup, G. and M. Latapy, Bipartite Graphs as Models of Complex Networks, in Combinatorial and Algorithmic Aspects of Networking. Springer. p. 127-139, 2005

[24] Shi, J. and J. Malik, Normalized cuts and image segmentation. Transactions on Pattern Analysis and Machine Intelligence, v.22, n.8, p.888-905. 2000.