A FRAMEWORK FOR A SOCIAL NETWORK OF RESEARCHERS ANALYSIS

Luciano A. Digiampietri[1], Ernando E. da Silva[2]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (EACH-USP)
Av. Arlindo Bettio, 1000, CEP 03828-000 São Paulo – SP – Brazil.
E-mail: digiampietri@usp.br

[2]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (EACH-USP)
E-mail: ernando@usp.br

Abstract. *Nowadays, a huge amount of data is available on the Web. One of the big challenges for Computer Science is how to help to select, organize and summarize this data. This paper presents an initiative in this direction in order to help researchers and commissions or foundations for research to access relevant information about researchers' groups. The approach developed combines Social Network Analysis, Graph Theory and Knowledge Discovery techniques to identify, organize, manage, visualize and summarize information about a Social Network of Researchers.*

Keywords. *Social network. Researchers' network. Data mining.*

1. Introduction

Nowadays, it is possible to find a substantial amount of data on the Web related to the most varied topics. One of the great challenges that Computer Science faces is how to extract, organize, summarize and visualize relevant information from this data.

Among these data there is important information for researchers, such as scientific publications, research projects information as well other researchers' curricula. When considering research data, Brazil has a peculiar feature: a national curricula database – the Lattes platform from the National Council for Scientific and Technology Development (CNPq), which gathers information about publications, academic advice, research projects, interesting areas, etc.

This vast and unique set of information is being underused. It is typically used to evaluate (or verify data) of research individually, or even as a data source for the manual verification of research groups and graduate programs.

This paper combines approaches from social network analysis, knowledge discovery and graphic theory in order to analyze the relationships among research groups and graduate programs, extracting, organizing and providing visualization tools for relevant data.

In this paper, social networks are created from researchers' curricula and these networks can be managed by the user and are characterized according to several metrics established by graph theory studies.

In addition, the rules and metrics adopted by the Computer Science Committees for the Graduate Program of the Brazilian Coordination and Improvement of Personnel with University Degrees (CAPES) for the evaluation of graduate programs were incorporated into the system to automate (at least partially) the individual and global analysis of the curricula of the researchers of a given graduate program in relation to those metrics.

The goal of this paper is to identify and analyze social networks of researchers who have a curriculum in the Lattes platform. In order to do this, information from the curricula will be used to establish relationships between researchers and thus enable the generation of a social network. Multiple metrics will be calculated from the produced networks (e.g., cohesion and centrality) to characterize them. Furthermore, a visualization tool was developed to facilitate the user to visualize and interact with the social network. Finally, a summarization tool was developed to help the analysis of the scientific production of a group of researchers based on metrics established by the Computer Science Committee of CAPES.

The rest of this paper is organized as follows. Section 2 presents some basic concepts. Section 3 details the related work. Section 4 contains the description of the proposed and developed framework for Social Networks of Researchers Analysis. Section 5 finishes the paper with some conclusions and suggestions for future work.

2. Basic Concepts

This paper deals with the problems of extracting data from the Web in order to represent social networks of researchers and facilitate the analyses of these networks according to some criteria. This section presents the basic concepts used in this paper, which include: social networks and graphs (Section 2.1); metrics for social network analysis (Section 2.2); the Lattes platform (Section 2.3); and CAPES and the Computer Science Committee (Section 2.4).

2.1 Social Networks and Graphs

Social networking is a social structure composed of individuals or organizations which are called "nodes" connected by one or more types of relationships of interdependence, for example, friendship, belief or work. There are several works developed in the area of social network analysis. Among them, we highlight those related to the area of exact sciences: network theory and graph theory [2,3,16], which established several metrics to analyze different characteristics of a social network.

One of the most common methods of computational representation of social networks is the use of graphs. Graphs are data structures that have been widely studied by computer science and applied to represent problems of several domains. A graph consists of a set V of vertices (or nodes) and a set E of edges. Each edge connects two vertices. A graph may or may not be directed. In a directed graph (also called digraph), the edges (also called arrows) have a direction, which means that the edge parts from the node A and reaches the node B (and this edge would be different from one which parts from the node B and reaches the node A). In an undirected graph there is no order relation between the nodes connected by an edge [5].

One of the ways to use graphs to represent social networks is to consider that each individual within a network is a node and each link (or relation) between individuals is an edge. The decision whether or not the graph is directed depends on the type of relationships between individuals that one wants to represent. For instance, the relationship of co-authorship does not require the use of directed edges: if the researcher A is the co-author of a paper with researcher B then, necessarily, B is also co-

author of a paper with researcher *A*. On the other hand, the advising relationship (advisor / student) is a relationship that requires a direction: if the researcher *A* is the advisor of the student *B*, it does not mean that the student *B* is the advisor of the researcher *A*.

There are several graph-related concepts that can be useful for the study and analysis of social networks. . These concepts are based on definitions from the work of [6], [11] and [15].

The focus on specific parts of a graph leads to the concept of a *subgraph*. A *subgraph* is a specific group of nodes and edges from the original graph. A *connected component* of a graph is a *subgraph* where all nodes can reach each other.

There is a difference in the notion of *node degree*, which depends on whether the graph is directed or undirected. In an undirected graph, the degree of a node is the number of edges connected to it. In a directed graph, this concept is divided into two: *indegree* and *outdegree*. The degree of a node can vary from 0 (when the node is isolated) to g-1 meaning the node is connected to all nodes in the graph (where g is the number of nodes in the graph) disregarding possible self-loops.

Some equations are now presented to describe some graph concepts. In these equations, *V* means the set of vertices of the graph; *E* is the set of edges; d(v) is degree of the node v; |V| is the total number of nodes; and |E| is the total number of edges. In this section, only the metrics associated with undirected graphs will be presented. For each of these concepts there is an equivalent in directed graphs.

The mean nodal degree corresponds to the mean of the node degrees of all nodes and it is given for the following equation:

$$\overline{d} = \frac{\sum_{v \in V} d(v)}{|V|}$$

The distribution of the nodal degree in a graph can be calculated using the degree variance (S_d^2):

$$S_d^2 = \frac{\sum_{v \in V} \left(d(v) - \bar{d}\right)^2}{|V|}$$

The *density* (Δ) is an important measure that can be used to compare different graphs. Density value ranges from 0 to 1:

$$\Delta = \frac{|E|}{|V| * (|V| - 1)}$$

A *path* in a graph is a sequence of nodes such that for all two consecutive nodes there is an edge linking them. If the same node starts and finishes the path, this path is called a *cycle* [8]. The size of a path is given by the number of edges of the path and it is equal to the number of the nodes of a path minus one.

The *range* is a concept related to the existence of a path between two nodes. If this occurs, these nodes are said to reach each other. If the *reachability* permeates all nodes in a graph, then the graph is considered *connected*.

The *geodesic distance* is a definition of distance between nodes which refers to the shortest path between two nodes of the graph. The *diameter of a graph* is defined as the largest geodesic distance in a connected graph.

2.2 Metrics for Social Network Analysis

According to [14], the choice and the application of metrics to specific issues require common sense from those who are conducting research on social networks. It is important to have a focus of analysis and select a set of metrics that may be relevant to a proper understanding of the network.

Considering the modeling of social networks through graphs, there are some metrics based on the graphs' properties, which are important for understanding the structure of the social network. Among these metrics, we highlight the centrality and prestige in a social network.

The *centrality* is related to the amount of connections of a node in relation to the entire network. The *prestige*, in turn, is related to the number of edges that reach a

particular node in a directed graph. Centrality and prestige can be calculated for a node or a set of nodes (subgraph) of a social network [14].

There isn't only one measure of centrality, but some definitions focus on specific concepts, such as the degree of centrality, the closeness centrality and the centrality of mediation, originally defined by Freeman et al. [7]. There are also the centrality of information and the centrality of status or rank [12,13].

The *normalized degree centrality* is calculated as follows [12]:

$$C_d(v) = \frac{d(v)}{|V|-1}$$

Typically, the individuals with greater centrality correspond to the more visible individuals in the network. They are connected to several nodes in the social network and, thus, are the most prominent [17].

The *closeness centrality* measures the proximity of one node in relation to the entire network [12]. This measure can be calculated only for connected graphs:

$$C_c(v) = \frac{\sum_{u \in V} dist(v,u)}{|V|-1}$$

Where *dist(v,u)* is the geodesic distance between the nodes v and u.

The closeness centrality is used to identify how far one node is from the center of the social network [12].

In order to verify the influence of one node in the paths between each pair of nodes in the social network the *betweenness centrality* was defined [7]:

$$C_b(v) = \sum_{x \in V} \sum_{y \in V}^{x \neq y} \frac{g_{jk}(v)}{g_{jk}}$$

Where g_{jk} is the number of geodesic paths between the nodes j and k; and $g_{jk}(v)$ is the number of geodesic paths between the nodes j and k and that have the node v in the path. The *betweenness centrality* can be normalized using the following equation [17]:

$$C'_b(v) = \frac{C_b(v)}{|V| * (|V| - 1)}$$

All these centralities were calculated for one node, but they can be calculated for a set of nodes in order to identify the influence of this group of nodes in relation to the entire network [13].

In [17] were adapted the presented metrics to deal with directed graphs. When considering directed graphs there are also others metrics that can be used such as *prestige* and *cohesion* [17].

2.3 Lattes Platform

The Lattes platform is provided by the National Council for Scientific and Technology Development (CNPq) and aims to be a data base for curricula for people that execute activities related to science and technology.

The curricula stored in the Lattes platform (also called Lattes curricula) are completed in the Web[1] and contain the following sections: personal information; academic degrees; research lines; research projects; interesting areas; knowledge of languages; titles and awards; scientific and technical production, among others. Each of these sections is divided into items that are filled in their specific forms. The filling of each item respect some rules but it is made manually by the research owner of the curriculum.

In order to allow the search for the curriculum of a given researcher the Lattes platform has a search service[2] but this service is only made for searches made by humans (there is a *captcha* security mechanism to avoid software agents using the service).

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

[2] http://buscatextual.cnpq.br/buscatextual

2.4 CAPES and the Computer Science Committee

CAPES is the Brazilian Coordination and Improvement of Personnel with University Degrees. It is part of the Brazilian Federal Education Ministry (MEC) and is responsible for authorizing and evaluating Brazilian graduate programs. In order to do so CAPES contains several committees, one for each scientific area. Each committee establishes a set of rules that will be used in the evaluation of the graduate programs that belong to this area.

Moreover, CAPES maintains a classification called Qualis[3] for evaluating scientific journals and conferences. Each committee establishes what are the relevant journals and conferences for its area and attributes a quality classification for this journal or conference. The classification can assume eight values: A1, A2, B1, B2, B3, B4, B5 and C, where A1 means the best classification and C is the worst one.

The Qualis classification is typically used for the evaluation of intellectual production of groups of researchers.

The Computer Science Committee attributed weights for each of the classification values, as in Table 1.

Table 1. Classification and weights attributed by the Computer Science Committee

Classification	A1	A2	B1	B2	B3	B4	B5	C
Weight	100	85	70	50	20	10	5	0

Moreover, this committee established two indexes to measure the intellectual production of researchers from a computer science graduate program: the general index and the specific index. The general index is given by the sum of the weights of all papers published by the selected researchers in the last three years divided by the number of researchers. And the specific index corresponds to the sum of the weights of only papers with classification A1, A2 or B1 published by the selected researchers in the last three years and divided by the number of researchers.

[3] http://qualis.capes.gov.br/webqualis

This criterion is used only to summarize the total intellectual production of a group of researchers. The evaluation rules established by the Computer Science Committee involve several other aspects such as collaborations, research projects, distribution of intellectual production, among others. The complete document is available at the CAPES website.[4]

3 Related Work

The literature on social network analysis is extensive and in the last sixty years several studies have been carried out [10]. The field for the research of social networks is very open, because it can exploit both the structure of these networks (in a closer approach of graph theory) and the relationships that it symbolizes in a given domain, exploring the social and cultural context of these relationships. This section describes only the related work that focuses on the analyses of social networks of researchers.

In [10] was investigated the structure of scientific collaboration networks, where relationships were based on the authors and co-authors of articles (two scientists are connected if they published an article together). The information used to build these networks was extracted from databases of biomedical, physics and computer science. The author shows that these networks are of a small-world type and that pairs of scientists are separated by short paths. The period considered for analysis was from 1995 to 1999 and possible tools used to extract the data, calculate metrics and display the resulting networks were not mentioned, which indicates that the work was largely performed manually. In her paper, Newman also reports that the major difficulty in this type of work is to estimate the correct number of authors in a database, since some authors may have the same name and may also call themselves differently in each paper (using the full name, only the first short name or all names with abbreviation).

Silva et al. [15] also conducted a study of a social network of research. The paper focuses on the network of co-authorship of professors from PPGC-UFMG (Graduate Program in Information Science at the Federal University of Minas Gerais).

[4] http://www.capes.gov.br

The aim of this work was to identify characteristics inherent in relationships between the actors involved in this network (network density, collaboration among different professors from different research lines). This work was based on the data from the Lattes platform. This paper analyzed the bibliographic production of professors from PPGC-UFMG during the period 1997 to 2004, undergoing a process of validation of data (such as standardization of the spelling of authors' names and classification of items according to the number of authors, for example) and using the UCINET [4] for the representation of the resulting network. The authors used the representation with both binary matrix (indicating whether or not an author published an article in conjunction with another) and matrix valued (which stores the number of articles published by the authors).

Other studies focusing on the extraction of data from the Lattes platform were developed by [9]. In their paper, the authors highlight the difficulty of obtaining and organizing data for medium and large groups of researchers from the Lattes platform since the public information of Lattes curricula is available only for research individually and requires a manual effort to collect data from groups of researchers. They describe the architecture, implementation and experiences with *scriptLattes*, an open source system. The *scriptLattes* takes as input a file containing a list of researchers, downloads these curricula for the subsequent generation of the network of collaboration and reporting. In this file, for each researcher there should be available the code (ID) Curriculum Lattes (which is used as reference for the page containing the curriculum of the researcher), the name and the period to be considered. This means that the user of the system should have prior knowledge about the researchers who will compose the reports and the network.

Another recent study on the extraction of data from the Lattes curriculum is being developed by [1] and is related to the creation of an API (*Application Programming Interface*) called *LattesMiner*. This tool intends to provide similar functionalities of the *scriptLattes*. This initiative is part of the project *SUCUPIRA* (Unified Curricula and Programs: Identifying Academic Networks) and has as its main goal the exploration of social networks research, but also to be a tool that allows the

extraction of scholarly information through the Web, to measure performance of researchers, professors and graduate programs. The *LattesMiner* tool is not available yet.

The framework presented in this paper differs from the related work for three main reasons. First, it contains search agents to find the Lattes curriculum of a given researcher in order to automate this task which has been typically done manually. Second, the framework provides a graphical environment to allow the user to manage and visualize the social network and obtain several metrics for each individual and for the entire network. Third, the reports produced by the system take into account the evaluation rules provides by the Computer Science Committee of CAPES established to evaluate groups of researchers from graduate programs.

4. The Framework

The developed framework is composed of 4 sub-systems: (i) Finding Researcher's Curriculum; (ii) Curriculum Data Extraction (iii) Social Network Visualization; and (iv) Summary Report Generation. Figure 1 contains an outline of the relationship of these sub-systems in the framework.

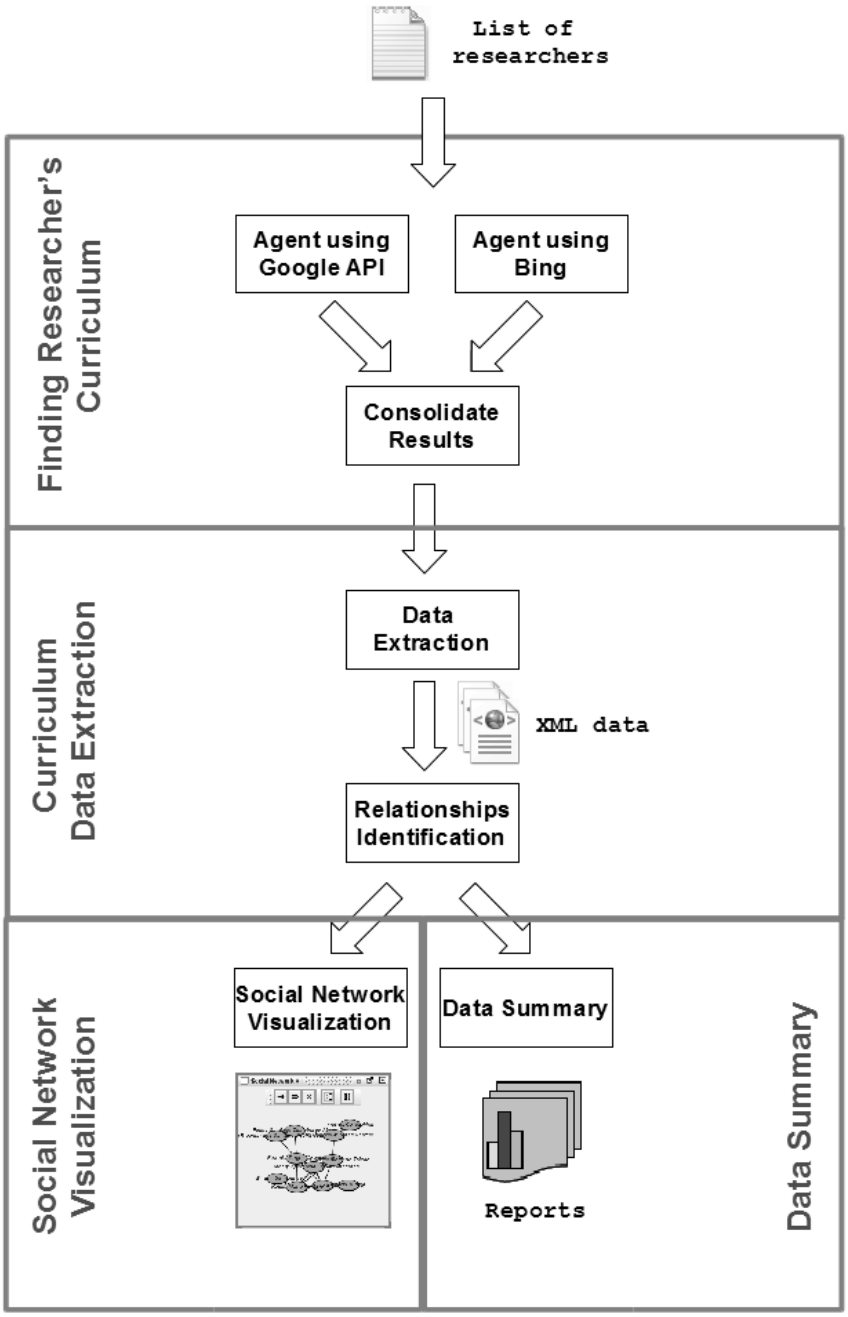

Figure 1. Framework overview

In the following sections each sub-system will be detailed. The framework was developed in Java and its current version is being tested and documented in order to be available on the Internet for anyone who wants to use or extend it.

4.1 Finding a Researcher's Curriculum

In order to find the URL of the researcher's curriculum two search information software agents were developed. The first uses the Google Search API and the second uses the Bing Web Search. These software agents execute a query in the Web using as keyword the researcher's full name and the expression '*currículo lattes*'. Each agent mines its Web search result to identify the URL of the searched curricula. Whenever a URL from a curriculum is identified the curriculum is downloaded and a validation is made to confirm if the curriculum belongs to the researcher of the search.

The system tested 1002 researchers' names and the results were verified manually to validate it. The test data for this validation is composed of researchers from 10 graduate programs in Brazilian universities and 240 professors from the Schools of Arts, Sciences and Humanities at the University of São Paulo. The full names of the researchers were obtained from the universities' websites. All the researchers have a curriculum in the Lattes System. Table 2 contains the results of the execution of this test considering 1002 researchers.

The number of true positives are the curricula identified correctly (belonging to the researcher that was being searched for). The number of false positives corresponds to the curricula that were wrongly identified as belonging to a given researcher. The number of true negatives are the curricula that were downloaded and verified and did not belong to the researcher. The number of false negatives means the curricula that were downloaded and, in the verification, were incorrectly identified as not belonging to the researcher whose curriculum was being searched for. From the 1002 curricula searched, the system found and downloaded 974 different curricula where 822 were correctly identified as belonging to the researcher of the search.

Table 2. Performance evaluation of the Finding Curricula system

Metric	Curricula Identified	Percentage
True positives	822	82%
False positives	0	0%
True negatives	136	-

False negatives	16	1,6%

4.2 Curriculum Data Extraction

The Researcher Data Extraction system is responsible for extracting the data from the HTML curriculum (semi-structured data) and converting it to XML data that will be used by the other sub-systems.

The current version of this system extracts information about intellectual production and interesting areas of each research.

The second function of this system is to identify the relationships among the researchers. Two kinds of relationships are being identified: the co-authoring relationship and the sharing of interesting areas relationship.

The interesting areas in the curricula in the Lattes System are selected from a list of pre-defined (sub-) areas. Thus, there is no risk of misspelling or any other kind of problem related to incorrect filling of a field. In order to identify researchers that have the same interesting areas the system must only match the list of interesting areas of each two researchers.

On the other hand, the identification of co-authoring relationships presents some challenges. The bibliographic information about each publication is filled in manually by the researcher using a form, which is specific for each kind of publication. The system must deal with some problems such as missing information and misspelling. To deal with these problems this system uses a basic text mining algorithm to verify if two publications (for different researchers) are the same. This algorithm verifies the edition distance of the titles and events (or journals) of the publication. Moreover, there is a specific algorithm for some kinds of publications. For example, for journal publications, the text mining algorithm is used to identify the ISSN of the journal (from a database of journal titles, abbreviations and ISSN); then the second step is, given two papers that were published in the same journal (with the same ISSN) the system will verify the titles of the papers and the number of the journal, the volume and the page numbers of the papers.

The data extracted from the curricula (including the relationships) are used for the Social Network Visualizations and Network Summary Report systems.

4.3 Social Network Visualization

In order to allow the user to visualize and edit the networks identified by the presented systems a graph management system was developed. This system can manage different kinds of graphs independently of the application domain and the measure of several graph metrics (see Section 2) was implemented in order to provide the user with relevant information about the graph (which in this application domain represents a social network).

Figure 2 contains a screenshot of the visualization tool. In this screenshot it is possible to see the co-authoring relationship between the years 2000 and 2010 among the professors from the Schools of Arts, Sciences and Humanities from the University of Sao Paulo. Only the professors who share some publications with other professors in this group are shown (no node without an edge is presented).

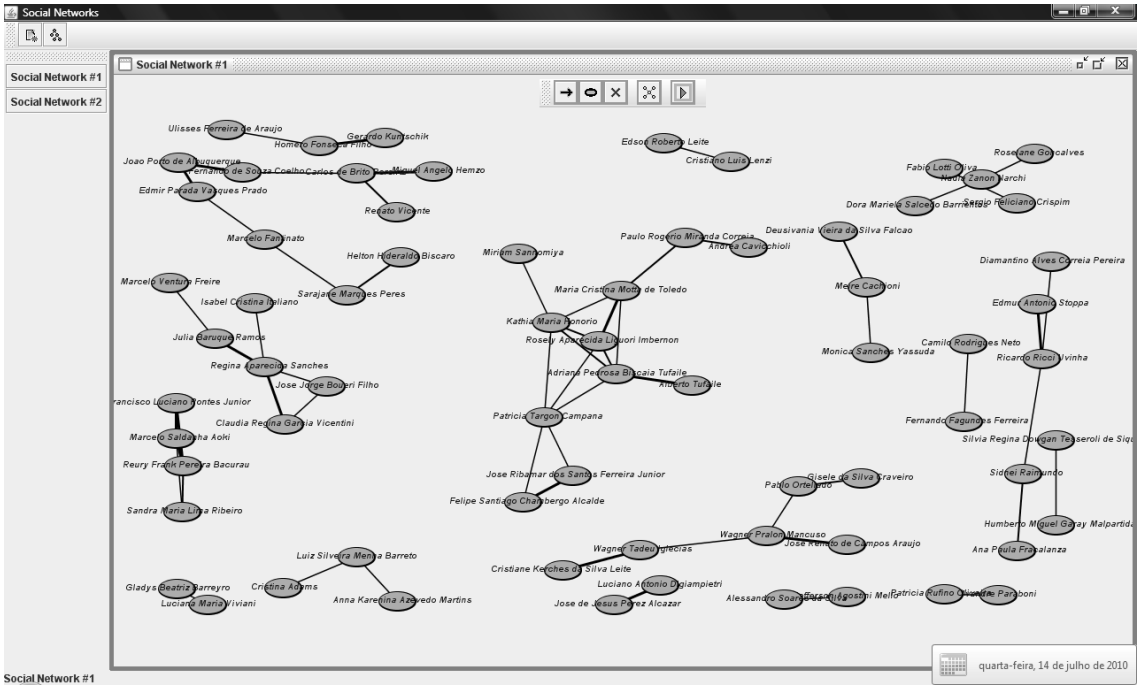

Figure 2. Screenshot of the Social Network Visualization tool

Figure 3 highlights a connected component (a subgraph) of the social network presented in Figure 2. The thickness of the edge is proportional to the amount of

publications that the two professors published together. In this figure it is possible to observe that *Adriana Pedrosa Biscaia Tufaile* published more papers with *Alberto Tufaile* than with *Patricia Targon Campana*.

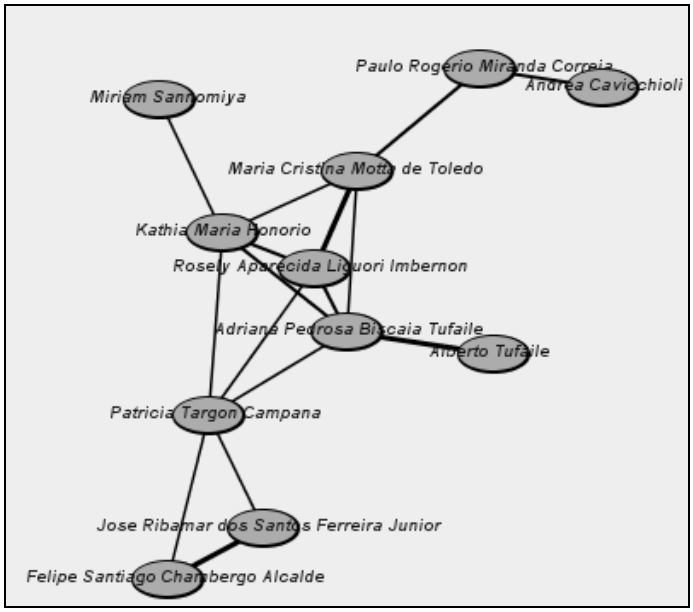

Figure 3. Subgraph of the Social Network presented in Figure 2

One of the functionalities of the visualization tool is to show the distance between the selected node (researcher) and its neighborhood. Figure 4 presents the subgraph of Figure 3 with the distance between each research and the researcher *Rosely Aparecida Liguori Imbernon.* The colors of each node correspond to the distance: the darker nodes are the nodes more distant.

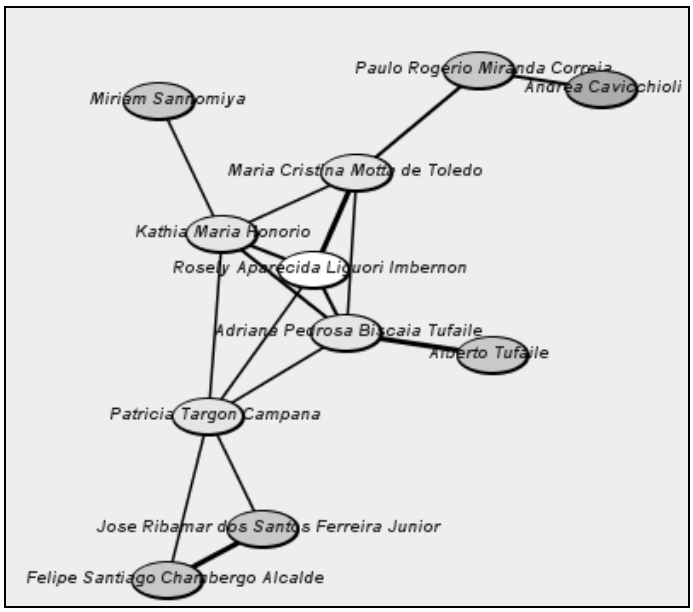

Figure 4. Distance between the central node and the rest of the network

The visualization tool has seven buttons in its graphical interface. Figure 5 presents these buttons and their names.

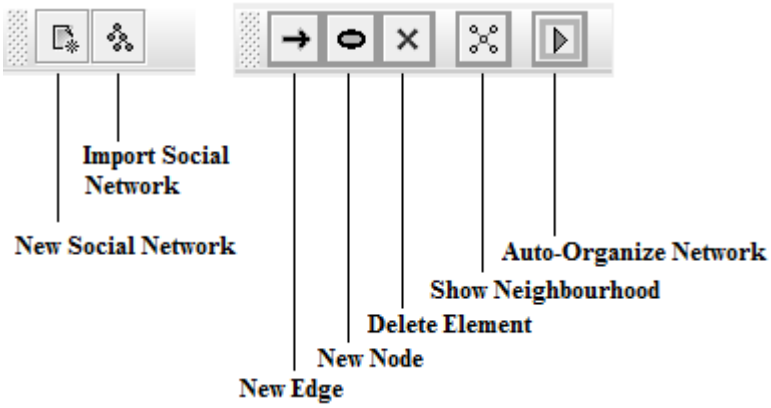

Figure 5. Buttons of the visualization tool

The *New Social Network* button opens a new empty social network. The *Import Social Network* button allows the user to select a text file with the names of the individuals of a social network and the relationships among those individuals. The *New Edge* button adds a new edge in the social network. The *New Node* button adds a new individual to the social network. The *Delete Element* button deletes the selected node or edge. The *Show Neighbourhood* button changes the color of the nodes according to their distance from the selected node. The *Auto-Organize Network* button iteratively organizes the graphical representation of the social network in order to facilitate its

visualization. The user can also move the nodes in the social network and whenever a user double clicks a node a form is opened with some metrics of the selected node and the entire graph. Figure 6 presents this form when the node *Kathia Maria Honorio* from Figure 1 is double clicked.

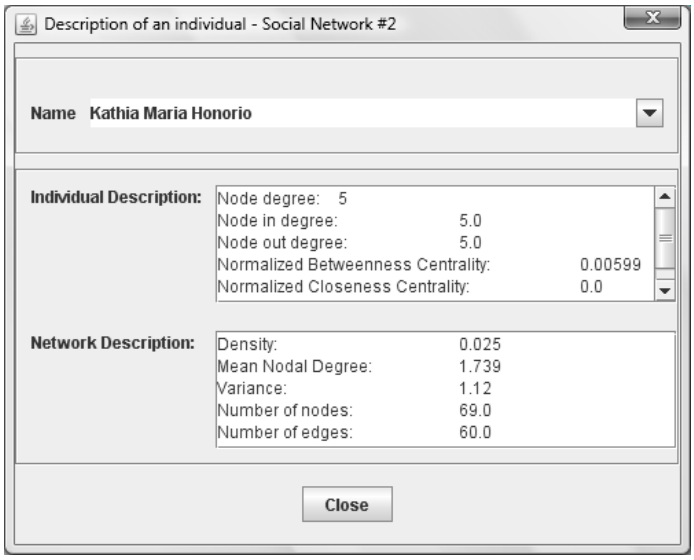

Figure 6. The description dialog for an individual in the network

With the same set of researchers presented in Figure 2 it is possible to generate other graphs in order to express different kinds of relationships. Figure 7 presents the graph of Social Network of Researchers where an edge means that two researchers share one or more interesting sub-areas.

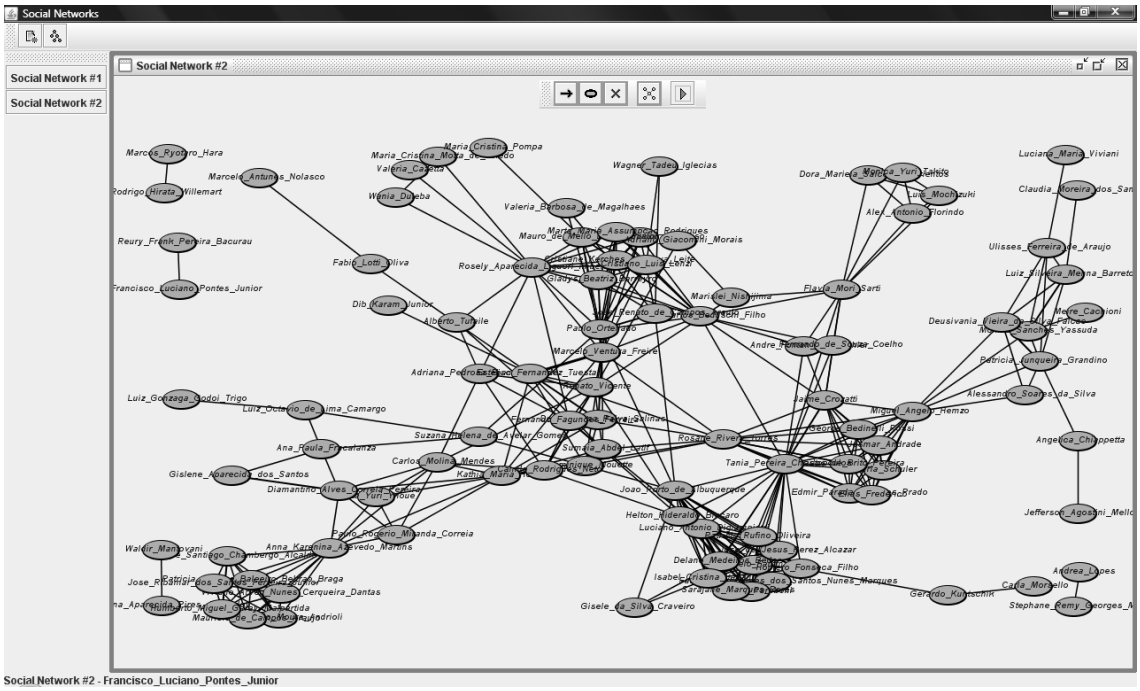

Figure 7. Screenshot of the Social Network Visualization tool – interesting areas

4.4 Summary Report Generation

Different interesting information can be generated from the data of a set of curricula, for example, the total number of published papers per year, the number of collaborations, the number of papers per researcher, etc.

The system of report summarization generation aims to complement the networks' information to facilitate understanding and evaluation of all researchers. Specifically, the system will analyze the intellectual output of researchers according to criteria found in the documents of the areas of CAPES. Since each area contains specific criteria, this system was developed to address only the rules of the Computer Science Committee.

As presented in Section 2.1, CAPES has a classification system for papers published in scientific journals and conferences, and the Area Committee rules uses this classification to evaluate the intellectual production of the set of researchers that belong to one graduate program. Thus, the first task of the summary generation system is to identify the classification of each paper. To do this, the system loads the list of all

classified journals from CAPES. Since the journals are indexed by their ISSN, it is necessary only to match the ISSN of a journal paper with the ISSN of the classified journals. For each journal paper the Data Extraction System had already identified its ISSN (Section 4.2). It is important to notice two things: not all journals are classified by each Area Committee from CAPES, and sometimes it is not easy to identify the ISSN of a journal paper in the Lattes Curriculum because this field is not mandatory and the journal name is manually filled in by the user, so it can contain spelling errors.

Conferences are classified only for the Computer Science Committee, thus to search for the classification of conference papers makes sense only for this area. Since conferences do not have a unique identifier (such as an ISSN) the matching between the classified conferences from CAPES and the conference of a paper from *Curriculum Lattes* is made considering the name and the abbreviation of the conference. This matching uses an edition distance algorithm and its efficiency depends on the correct spelling of the conference name. It is important to remember that the Computer Science Committee associated a numeric value to each classification, so it is possible to do numerical analyses considering the "quality" of each paper.

Once the classification of the journal and conference papers is made, the system counts the amount of papers belonging to each class and produces a report grouping this data by author, year and class. Moreover, the system calculates the general index and the specific index of the group of researchers that are being analyzed.

Figure 8 presents the summary of publication per researcher from a M.A. graduate program at the University of São Paulo for the years 2000 to 2010. Figure 9 presents the summary of the same researchers for the years 2008, 2009 and 2010.

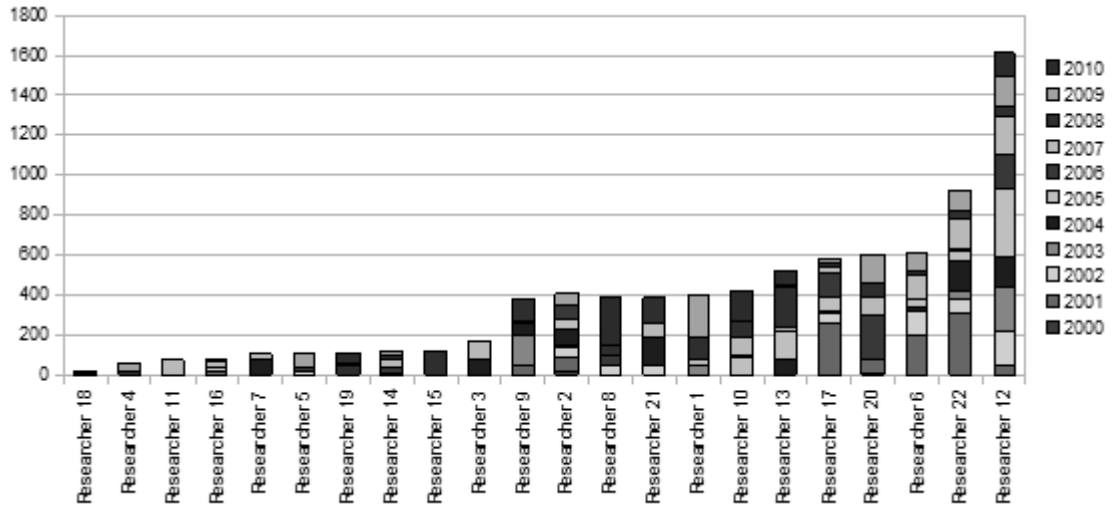

Figure 8. Summary of publications for a group of researchers – 2000 to 2010

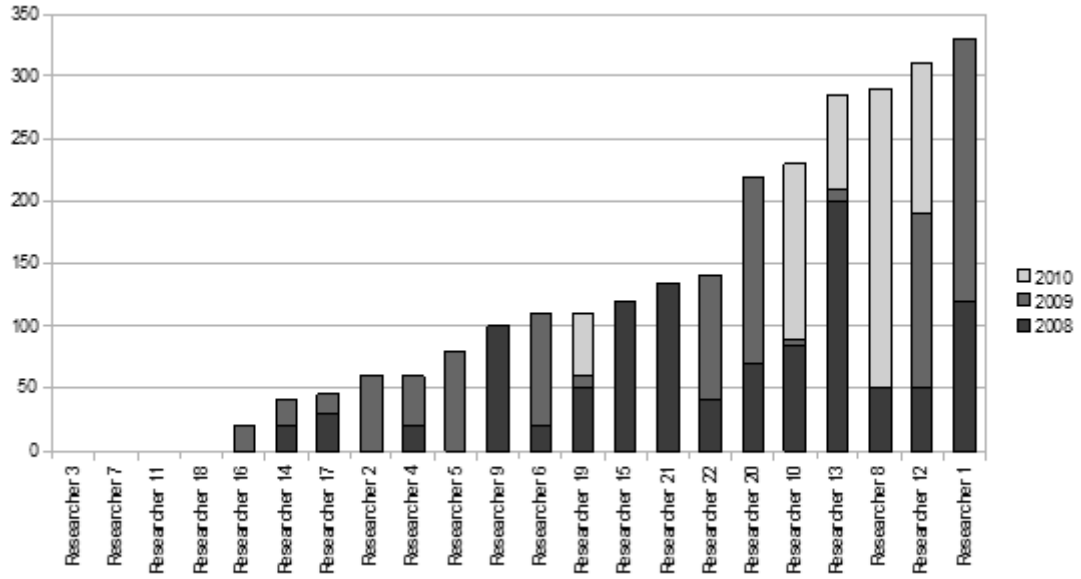

Figure 9. Summary of publications for a group of researchers – 2008 to 2010

The summary report also contains the general and specific indexes defined by the Computer Science Committee. Table 3 presents the value of these indexes for the researchers on Figure 9.

Table 3. Indexes for the period 2008-2010

General index	1.82
Specific index	0.93

The report also contains information about how many papers were published per year and how many are collaborations among two or more researchers of the group being analyzed.

4. Conclusion and Future Work

Every society is characterized by different kinds of relationships between the individuals who compose it. It makes sense that one tries to study, visualize and characterize the various existing social networks. This type of study is known as social network analysis and appears as a focus of research on several occasions in the literature. It is much more than a strategy for research in social structures than a formal theory [13].

This paper presented a framework for the analysis and visualization of Social Networks of Researchers. The framework executes a full process of knowledge discovery starting with the search for the curricula of the relevant researchers; the information extraction of these curricula; and relationships identification. With this information the social network is produced and it can be manipulated using a graphical tool and a summary of the intellectual production is produced according to the criteria defined by the Computer Science Committee of CAPES.

This work extends related work in three aspects. First, it searches the Web in order to find the Curriculum Lattes of a given researcher. Second, it contains a graphical tool that allows a user to manipulate the social network and to obtain metrics about the individuals and the entire network. Since these metrics are generic graph-based metrics they can be used in any application domain (not only in researchers' networks). Third, the summarization report contains information about the Social Network of Researchers considering the rules defined by the Computer Science Committee of CAPES in order to combine the researchers' intellectual production with the classification of journals and conference papers and the other criteria of graduate programs evaluation.

The analysis of researchers' networks can be used in order to identify possible collaborations, interesting groups or even experts in a given area.

The future work related with this paper includes: (i) increase the efficiency of the algorithm for finding the researcher Lattes curriculum; (ii) increase the efficiency of the algorithms that matches the titles of journals and conferences with those named in the publications of each Curriculum Lattes; (iii) extends the summary generation system to produce reports considering other committees' rules and not only the Computer Science Committee. Other possible future extension of this work involves the development of a framework to relate scientific publication in order to guide users in search of the bibliography about a desired subject.

References

[1] A. D. Alves, H. H. Yanasse, N. H. Soma (2009). Extração de Informação na Plataforma Lattes para Identificação de Redes Sociais Acadêmicas. *Proceedings of the IX Workshop do Curso de Computação Aplicada - Poster Track*. São José dos Campos, Brazil.

[2] S. D. Berkowitz (1982) An Introduction to Structural Analysis: The Network Approach to Social Research. Butterworths.

[3] R. L. Breiger (2004). The Analysis of Social Networks. Handbook of Data Analysis. London, Sage Publications.

[4] S. P. Borgatti, M. G. Everett, L. C. Freeman (2006) UCINET 6 for Windows - software for social network analysis: User's guide. *Analytic Technologies.*

[5] T. Cormen, C. E. Leserson, R. L. Rivest (2001). Introduction to Algorithms.The MIT Press. Second Edition.

[6] R. Diestel (2006). Graph Theory. Springer-Verlag, Third edition.

[7] L. C. Freeman. Centrality in social networks: Conceptual clarification. Social Networks, 1(3), pages 215-239.

[8] R. A. Hanneman, M. Riddle. (2005). Introduction to social network methods. Riverside, California.

[9] J. P. Mena-Chalco, R. M. Cesar Junior (2009). scriptLattes: an open-source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society*, 15(4):31-39.

[10] M. E. J. Newman (2001). The structure of scientific collaboration networks. *Proceedings of The National Academy of Sciences of USA (PNAS)*, volume 98, pages 404-409.

[11] M. E. J. Newman (2003). The structure and function of complex networks. *SIAM Review*, 45(2), pages167-256.

[12] E. Otte, R. Rousseau (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6), pages 441-453.

[13] D. A. Población, R. Mugnaini, L. M. S. V. C. Ramos (2009). Redes sociais e colaborativas em informação científica. Editora Angellara, São Paulo.

[14] J. Scott (2000). Social network analysis: a handbook. Sage Publications. Second edition.

[15] A. B. O. Silva et al. (2006). Estudo da Rede de Co-Autoria e da Interdisciplinaridade na Produção Científica com Base nos Métodos de Análise de Redes Sociais: Avaliação do Caso do Programa de Pós-Graduação em Ciência da Informação – PPGCI-UFMG. Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação, volume 10, 16 pages.

[16] B. Ulrik, T. Erlebach (2005). Network Analysis: Methodological Foundations, Heidelberg: Springer-Verlag.

[17] S. Wasserman, K. Faust (1998). Social network analysis: methods and applications. Cambridge University Press.