11th World Congress on Computational Mechanics (WCCM XI)
5th European Conference on Computational Mechanics (ECCM V)
6th European Conference on Computational Fluid Dynamics (ECFD VI)
E. Oñate, J. Oliver and A. Huerta(Eds)

AN EFFICIENT PROCESS FOR EXTRACTION AND IDENTIFICATION IN SCIENTIFIC COLLABORATION NETWORKS

THIAGO M. R. DIAS AND GRAY F. MOITA

Programa de Pós-Graduação em Modelagem Matemática e Computacional (PPGMMC)
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
Av. Amazonas, 7675, CEP 30510-000, Belo Horizonte, MG, Brazil
e-mail: thiagomagela@gmail.com and gray@dppg.cefetmg.br

Key Words: *Extraction and data integration, Information Retrieval, Scientific Collaboration.*

Abstract. The analysis of scientific collaboration networks has significantly contributed to improving the understanding of how does the process of collaboration between researchers occurs and also to identify how the evolution of scientific production of researchers or research groups can be understood. However, the identification of collaborations in large scientific databases is not a trivial task given the high computational cost of the methods commonly used. This paper proposes a method for identifying collaboration in large database of researchers´ curriculum. The proposed method has a low computational cost, with quite satisfactory results, proving to be an interesting alternative for the modeling and characterization of large scientific collaboration networks.

1 INTRODUCTION

Lately, in addition to scientific production, there has been a steady growth in the study of networks in relation to various disciplines ranging from computer science and communications to sociology and epidemiology. A network can be characterized as a graph that consists of a set of nodes (vertices) and links (edges) between the nodes. These links can be either directed or not directed, and can optionally have an associated weight. Many, perhaps most, natural phenomena can naturally be described in terms of a network. The brain may be characterized as a network of neurons connected by synapses. The Internet is also an example of an important network for society today.

The subject has been studied by several researchers; nonetheless, it was only recently that the analysis of networks has become an important area of research. This is partly due to the advancement of computers. Computers have aided in the empirical study of real networks, and have enabled researchers from different fields to conduct technical analyses of large networks.

The strong relationship between the scientific and the socio-economic domain has led to a growing interest in understanding the mechanisms involved in scientific activities. Furthermore, it has resulted in many studies that analyze the elements of its construction and

the characteristics of language and discourse used in scientific communication. The ratio of collaboration between researchers has also been analyzed [1].

In the scientific domain, an example of a social network can be considered as the scientific collaboration network that is observed as a graph in which the vertices correspond to the authors of specific scientific publications and the lines or edges correspond to co-authorship relationship. In this type of network, the mayor edges might not be weighted. The addition of weight would represent the number of joint papers in which the authors connected by the edge under analysis have participated.

For Stroele et al. [2], scientific social networks are specific types of social networks that represent social interactions originating in the academic environment. These interactions usually occur through the publication of scientific articles, academic guidelines and the development of research projects. Various activities and purposes can lead to the study of scientific collaboration networks, such as recommendation of new collaborators, intensifying the collaboration, ranking of groups or individuals, or identifying groups and their characteristics.

From the data available in scientific publications, it is possible to build collaborative networks. This work presents an efficient process for identifying collaborations in large scientific databases for the modelling and characterisation of networks wherein the vertices represent the authors and the edges represent papers co-produced by two or more authors.

2 RELATED WORK

With increasingly fierce competition among organizations and research institutions, it becomes important for members to discover potential collaborators in order to leverage the scientific production. Recent studies show that research groups with a well-connected social network science tend to be more productive [3,4].

Networks of co-authorship of a community can reveal interesting facts about them such as which groups collaborate better, the intensity of relationships between authors, or which authors work with a greater degree of collaboration. The study of networks of co-authorship can also be used to compare the patterns of collaboration between different scientific communities [5].

Canibano and Bozeman [6] have suggested that the curriculum vitae method can be used as a sufficiently comprehensive source of information in academic research, and that its usefulness has been widely explored since the beginning of 2000. However, few studies have investigated the use of curricula for conducting social network analysis, whereas several others have analyzed co-authorship and the impacts of scientific collaboration on the career of a given researcher [7].

The study by Petersen et al. [8] highlights factors that are of great importance to academic success in scientific networks. These factors include the abundance of scientific literature that enhances the attractiveness and the size of future opportunities for employees, and the co-author collaboration network. In view of this, it is evident that further study is necessary in order to understand and analyze how scientific collaboration happens as well as to design new tools aimed at boosting scientific production.

Other suggestions and applications for social network analysis research can be seen in [9-14]. These are based on the potential for mining, visualization and the structural analysis of

social networks of researchers, institutions or groups, and the thematic research in a particular area, based on their scientific productions — especially scholarly articles produced by the researchers.

Identification of contributions is a complex task mainly due to the nature of the data to be analyzed. This data usually does not have a well-defined pattern, introduces misspellings and lacks uniformity in the various ways in which an author can cite a collaborator in his work. Therefore, an efficient method to perform the identification of scientific collaborations, particularly in large databases, is required.

3 DEVELOPMENT

Data from the CNPq Lattes Platform were used to produce this work. The Lattes Platform was conceived to integrate the information systems of Brazilian federal agencies, optimizing the Science and Technology (S&T) management process from the standpoint of the user, the sponsorship agencies and the institutions of education and research. [15]

The selection of the Lattes Platform for extraction is related to the fact that it is extremely rich. This is because it deals with the integration of scientific data, both curricular and institutional in the S&T field, recording academic data and scientific production from researchers and institutions, allowing the researchers themselves to update the information. Currently, the Lattes Platform includes 3.5 million recorded curricula vitae.

Several papers for scientific data analysis have explored the Lattes Platform as a primary source of information. In Dias et al. [15], the whole process of extraction and data integration is divided into three main parts: Extraction, Processing, and Visualization (Figure 1). However, for the purpose of the current study, only the results of the extraction step that have the details of the curriculum and the subject of the research study were used.

The data extraction process in the framework begins with the acquisition of identifiers for the Lattes curricula that have been obtained with a requisition on the Platform. These identifiers are then stored locally. The acquisition strategy begins with a request that results in a list containing all the identification codes of the registered curricula.

Subsequently, a crawler collects the identifiers and generates a list of codes that will allow access to the curriculum of each individual researcher for extraction. All the extracted Lattes curricula are stored in the eXtensible Markup language (XML) format.

Each title in a given publication undergoes a transformation designed to obtain the same title without words with any semantic value, without any kind of accent marks and without spaces. Next, the entire text is standardized in lowercase characters and the resulting string is transformed into a key. An example of the process is shown in Table 1.

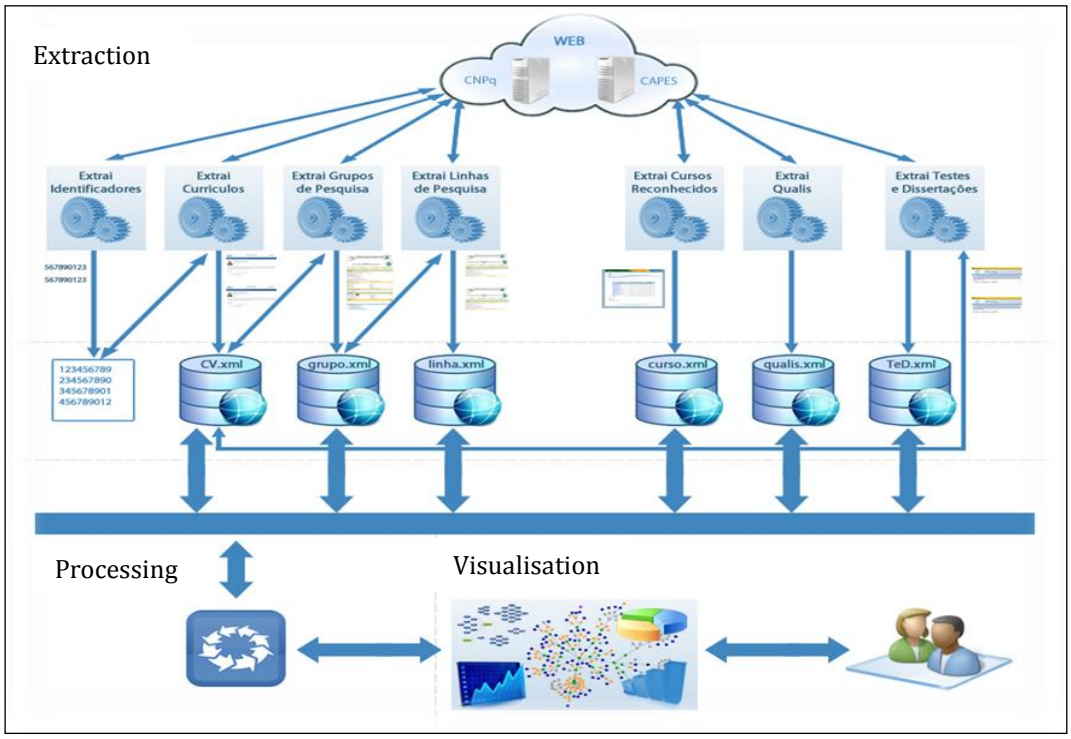

Figure 1: Framework for Extraction Lattes Platform. Dias et al. 2013.

Table 1: Example of the transformation process used

Steps	Result
1	Modeling and Characterization of Scientific Networks: A Study of the Lattes Platform 2013
2	Modeling Characterization Scientific Networks: A Study the Lattes Platform 2013
3	ModelingCharacterizationScientificNetworksAStudytheLattesPlatform2013
4	modelingcharacterizationscientificnetworksastudythelattesplatform2013

As a result of this transformation stage, a comparison is performed to verify whether this key already exists in the directory used to identify collaborations. Should the key already exist in the directory, the author's identifier in the curriculum vitae under analysis is inserted in the key position; otherwise, the key and the identifier are inserted in the abovementioned directory. Table 2 brings an example of the above construction.

Table 2: Example of the dictionary construction

Key	Author
modelingcharacterizationstudyscientificnetworkslattesplatform2013	Id01, Id25
studyaboutinfluenceacademicperformancestudentsuserssocialnetworks2013	Id25, Id145, Id98
analysiscollaborationnetworksscientificpublications2013	Id01, Id25, Id85
….	
….	
identificationprocessreviewersscientificnetworks2013	Id01, Id25, Id174

An example of the identification process of the proposed method can be seen in Table 3, where researchers, the true degree of collaboration (relationships) and the degree identified by the proposed method are presented.

Table 3: Collaboration identification process

Researcher_ID	Real collaboration degree	Value obtained by the proposed method
1	0	0
2	1	1
3	0	0
4	2	1
5	7	7
6	5	5
7	6	5
8	0	0
9	0	0
10	5	5
11	4	4
12	6	6
13	6	6
14	6	6
15	3	3
16	2	2
17	2	2
18	3	3
19	4	4
20	0	0
21	5	5
22	2	2
23	7	7
24	0	0
25	8	8

From the above, it is possible to observe that the method proposed in this paper has a very good level of precision, being able to identify a large number of relationships.

4 RESULTS

After analyzing all the article titles, a network is generated by building networks of cliques [16]. A set of vertices is called a clique when all the vertices are interconnected. Therefore, for each dictionary key that corresponds to an article, the elements representing the authors of the article are inserted into the network as a clique. Since a vertex to be inserted may already be present in the collaboration network, the clicks are linked by the juxtaposition of the common vertex (see Figure 2).

Importantly, using the proposed method in this paper, the construction of networks with a large number of publications can be modelled at a low computational cost and in an acceptable time, unlike the methods of cross validation where the computational cost is very high.

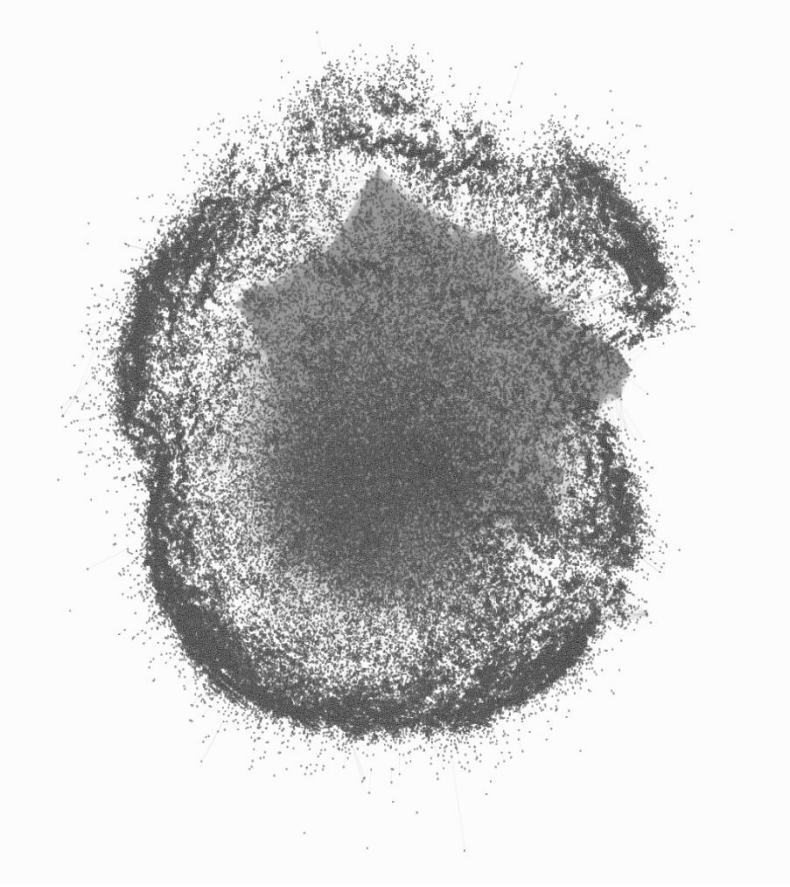

Figure 2: Network with 3,001,980 listed CVs (as in November 2013)

After modelling the network, it is possible to identify the vertices that have greater intensity of collaboration. These vertices are characterized by thicker edges. The vertices are then sorted according to research areas and several other features that can be drawn from the author information. It is possible to apply various metrics of social network analysis to a network constructed in this manner, in order to better understand the specific characteristics of each vertex as well as the topological characteristics of the network. Other examples of networks can be seen in Figure 3.

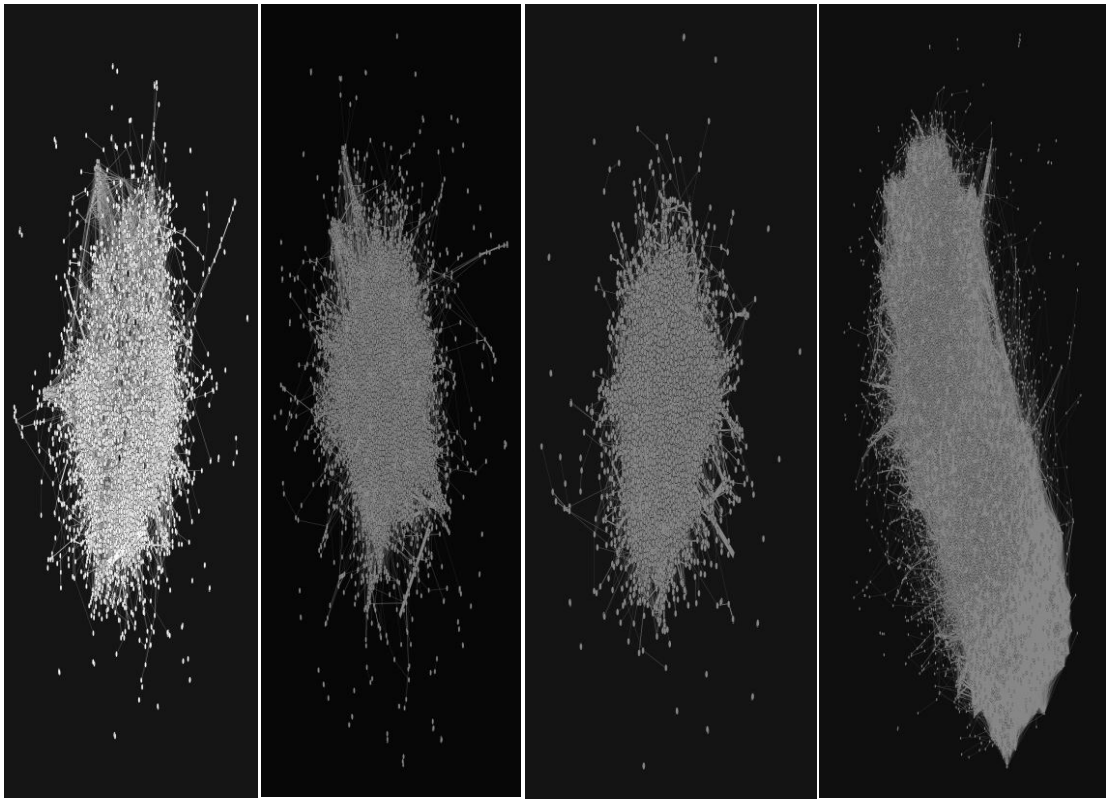

Figure 3: Example of networks obtained with the proposed method

5 CONCLUSIONS

In the present work, an efficient method for identifying collaboration in large database of researchers´ curriculum is proposed. The network is obtained by the juxtaposition of cliques on a collaboration graph.

The greatest advantage of adopting the current technique is with regards to its computational cost. As only one comparison for each article title is performed, it is possible to generate the collaboration network at a linear cost ($\theta(n)$), thus allowing the construction of the network in a much more interesting timeframe than with algorithms that work with cross-validation at a polynomial cost ($\theta(n^2)$).

6 ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support provided by the CEFET-MG, FAPEMIG and CAPES during this work.

REFERENCES

[1] Y. Ding, Y. Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. **J. Informetrics**, n. 5, v. 1, 2011, p. 187-203.

[2] V. Ströele, G. Zimbrão and J. M. Souza. Análise de redes sociais científicas: modelagem multi-relacional. In **Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)**, Curitiba, Brazil, 2012

[3] M. A. Brandão and M. M. Moro. Recomendação de Colaboração em Redes Sociais Acadêmicas Baseada na Afiliação dos Pesquisadores. In **SBBD - Simpósio Brasileiro de Bancos de Dados**, São Paulo, Brazil, 2012

[4] G. R. Lopes, M. M. Moro, R. da Silva, E. M. Barbosa and J. P. M. Oliveira. Ranking Strategy for Graduate Programs Evaluation. In **ICITA - 7th International Conference on Information Technology and Application**, Sydney, Australia, v. 1, 2011, p. 253-260.

[5] P. S. P. Júnior, A. H. F. Laender, M. M. and Moro. Analysis of Network Co-authoring the Brazilian Symposium on Databases. In **SBBD - Simpósio Brasileiro de Banco de Dados**, Florianópolis, Brazil, 2011.

[6] C. Cañibano and B. Bozeman. Curriculum vitae method in science policy and research evaluation: the state-of-the-art. **Research Evaluation**, n. 18, v. 2, 2009, p. 86-94.

[7] L. A. Digiampietri, R. Mugnaini and C. Alves. Analysis of Participation in supervised production of Advisors: A Case Study in Computer Science. In **Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)**, Maceio, Brazil, 2013.

[8] A. M. Petersen, M. Riccaboni, H. E. Stanley and F. Pammolli. Persistence and uncertainty in the academic career. **Proceedings of the National Academy of Sciences**, n. 109, v. 14, 2012, p. 5213-5218.

[9] Dias, T. M. R., & Moita, G. F. Extraction and Modeling of Scientific Collaboration Networks. In **Conferência IADIS Ibero-Americana WWW/Internet**, Porto Alegre, Brasil, 2013 (Vol. 1)

[10] V. P. Freire and D. R. Figueiredo. Ranking in collaboration networks using a group based metric. **Journal of the Brazilian Computer Society**, n. 17, v. 4, 2011, p. 255-266.

[11] J. P. Oliveira, G. R. Lopes and M. M. Moro. Academic Social Networks. **In International Conference on Conceptual Modeling (ER 2011)**, Brussels, Belgium, 2011.

[12] J. P. Mena-Chalco, C. Junior, C. and R. Marcondes. ScriptLattes: an open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, n. 15, v. 4, 2009, p. 31-39.

[13] H. A. Reijers, M. Song, H. Romero. U. Dayal, J. Eder and J. Koehler. A collaboration and productiveness analysis of the BPM community. In **Business Process Management**, Springer, 2009, p. 1-14.

[14] G. O. Fernandes, J. O. Sampaio and J. M. Souza. XMLattes - A Tool for Importing and Exporting Curricula Data. In **WORLDCOMP'11 - The 2011 World Congress in Computer Science, Computer Engineering and Applied Computing**, Las Vegas, Nevada, USA, 2011.

[15] T. M. R. Dias, G. F. Moita, P. M. Dias, T. Moreira and L. Santos. Modeling and Characterization of Scientific Networks: A Study of the Lattes Platform. In **Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)**, Maceió, Brazil, 2013.

[16] I. Fadigas and H. Pereira. A network approach based on cliques. **Physica A: Statistical Mechanics and its Applications**, n. 392, v. 10, 2013, p. 2576-2587.