Analyzing the Coauthorship Networks of Latin American Computer Science Research Groups

Juan F. Delgado-Garcia, Alberto H. F. Laender and Wagner Meira Jr.
Computer Science Department
Federal University of Minas Gerais
31270-901 - Belo Horizonte - Brazil
{*jfdgarcia, laender, meira*}*@dcc.ufmg.br*

***Abstract*—In this paper, we analyze the coauthorship networks of Latin American Computer Science research groups from 35 academic institutions in Argentina, Brazil, Chile, Colombia, Cuba, Mexico, Peru, Uruguay and Venezuela. Our analysis is based on data over a period of 20 years collected from DBLP, and aims to know the topological structure of each of these networks and provide a view of how they have evolved over time. Our results show that over the 2004-2013 decade there has been a relevant increase in terms of publications and collaborations in Latin America. We also identify the influential authors in the area according to complex network metrics and analyze the research networks originated from the coauthorships. Despite the increase in all per-country metrics, we observed that there is still a lot to improve, since most of the collaborations happen between just Brazil-Chile and Argentina-Brazil, although there is some growth in the diversity of the collaborations.**

***Keywords*-Latin American; Coautorship Networks; Bibliometrics;**

I. Introduction

According to Newman [1], a social network is a collection of individuals or groups of individuals connected by some kind of relationship that exist among them. Such individuals or groups are called *actors* and their relationships *ties*. This kind of network can be represented by a graph in which nodes denote actors and edges represent a specific tie among them.

A coauthorship network, is a special type of social network in which the actors represent authors and the ties indicate that the authors have coauthored at least one publication together. Due to the large amount of bibliographic data made available today on the Web, coauthorship networks have been widely studied over the past years [1], [2], [3], [4], [5], providing an interesting view of the academic communities behind them. Among the pioneering works, Newman [1] analyzes three scientific communities (Computer Science, Physics and Biomedicine) and presents several structural and topological features of them. Similarly, by mapping the publications from important journals in Mathematics and Neurosciences over a eight-year period (1991-1998), Barabási et al. [6] infer the dynamic and the structural mechanisms that govern the evolution and topology of the coauthiorship networks of these two communities based on several metrics.

In the context of the Computer Science (CS) area, Liu et al. [2] present a study of the Digital Library community based on the coauthorship network derived from its three main conferences. In such study, the authors analyze several aspects of this network, including its main connected components and the clustering coefficient. A similar study has also been carried out for the ACM SIGMOD Conference [5]. A comprehensive study of the CS area as whole based on data from CiteSeer[1] has been done by Huang et al. [3], whereas Menezes et al. [7] assess how the process of knowledge production in the area happens in three different geographic regions of the globe: Brazil, North America (Canada and US) and Europe (France, Great Britain and Switzerland).

Regarding specifically the Brazilian research community, Maia et al. [4] present a detailed analysis of the structural features and the evolution of the coautorship network of SBRC[2] thoughout its 30 years of history. In an attempt to contrast the scientific production of the Brazilian and international CS communities, Freire and Figueiredo [8] compare two coauthorship networks generated from DBLP: a global one, created considering all publications found in that digital library, and its subset that considers only researchers affiliated to Brazilian institutions. Finally, using data from the Brazilian National Research Council Lattes Platform[3], Mena-Chalco et al. [9] present a comprehensive study of the Brazilian scientific community by characterizing and exploring its main coauthorship networks. The study aims at gaining an in-depth understanding of the network structures as well as of the dynamics (social behavior) among the researchers in the eight major Brazilian knowledge areas: agricultural sciences, biological sciences, exact and earth sciences, humanities, applied social sciences, health sciences, engineering, and linguistics, letters and arts.

In this paper, we present an analysis of the coautorship networks of Latin American CS research groups. Our analysis is based on data over a period of 20 years (1994-2013) collected from DBLP [10], and addresses 35 institutions from Argentina, Brazil, Chile, Colombia, Cuba, Mexico, Peru, Uruguay and Venezuela. The main aim of

[1]http://citeseerx.ist.psu.edu/index
[2]Brazilian Symposium on Computer Networks and Distributed Systems
[3]http://lattes.cnpq.br

978-1-4799-6953-1/14 $31.00 © 2014 IEEE
DOI 10.1109/LAWeb.2014.13

our analysis is to know the topological structure of each one of these networks and provide a view of how they have evolved over time. Among our main results, we show that there has been a significant increase in the number of publications in the last decade as well as a consolidation of the research groups in some countries. We also identify the influential authors in the area according to three complex centrality network metrics. Finally, we analyze the research networks originated from the coauthorships.

The rest of this paper is organized as follows. Section II describes LACompNet - The Latin American Computer Science Network, a platform we have constructed to support our analysis, Section III discusses our main results, and Section IV presents our conclusions and summarizes future work.

II. LACompNet: The Latin American Computer Science Network

A coauthorship network can be represented as a graph $G_c = (V_c, E_c)$, where V_c represents a set of authors from a community c (e.g., an institution or a specific research group) and E_c represents a set of edges or coauthorship relations between two or more authors in V_c. In other words, an edge between two vertices indicates that the corresponding authors have coauthored at least one publication. In this context, **LACompNet** is a coauthorship network composed of Computer Science researchers from Latin American countries. Here, we consider LACompNet as a non-directed graph, where each vertex's weight corresponds to the number of publications per author during a period, and each edge's weight corresponds to the number of common publications between the corresponding authors.

Data Gathering. The data gathering process consisted of three steps: (i) determining the list of researchers (authors) of each institution of interest, (ii) collecting data from DBLP, and (iii) creating a relational database.

In the first step, we determined the list of researchers by manually extracting their names from the official websites of the 35 Latin American CS graduate programs considered. We then checked whether there was a DBLP entry for each of these researchers and collected their publication data from there. The current version of LACompNet covers researchers from the following institutions:

Argentina - Universidad de Buenos Aires (UBA), Universidad Nacional de la Plata (UNLP), Universidad Nacional del Centro de la Provincia de Buenos Aires (UNICEN) and Universidad Nacional del Sur (UNS); **Brasil -** Universidade Federal do Rio de Janeiro (UFRJ), Universidade Federal de Minas Gerais (UFMG), Universidade Federal do Rio Grande do Sul (UFRGS), Pontifícia Universidade Católica do Rio de Janeiro (PUC-RIO), Universidade Estadual de Campinas (UNICAMP), Universidade Federal de Pernambuco (UFPE), Universidade de São Paulo (USP) and Universidade de São Paulo at São Carlos (USP-SC); **Chile -** Pontificia Universidad Católica de Chile (PUC-Chile), Universidad de Chile (UCHILE), Universidad Santiago de Chile (USACH), Universidad Técnica Federico Santa Maria (UTFSM) and Universidad de Concepción (UDEC); **Colombia -** Universidad ICESI (ICESI), Pontificia Universidad Javeriana at Cali (PUJ-Cali), Universidad de los Andes (ANDES), Universidad del Valle (UNIVALLE) and Universidad Nacional de Colombia (UNAL); **Cuba -** Universidad de La Habana (UH), Universidad de las Ciencias Informáticas (UCI) and Universidad de Oriente (UO); **Mexico -** Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (CINVES-TAV), Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM), Universidad Autonoma del Estado de Mexico (UAEMEX) and Universidad Nacional Autonoma de Mexico (UNAM); **Peru -** Universidad Católica San Pablo (USCP); **Uruguay -** Universidad de la Republica (UDELAR); **Venezuela -** Universidad Central de Venezuela (UCV), Universidad de Carabobo (UC), Universidad Simón Bolívar (USB) and Universidad de Los Andes (ULA).

In the second step, we collected data from the DBLP entry of each identified researcher. In particular, we collected data from a 20-year period that ranges from 1994 to 2013. Finally, in the third step, we used a Java crawler to extract the data of interest from the DBLP pages. Then, we used the Simple API for XML (SAX[4]) for parsing the resulting XML files, and populated a relational database (MySQL) to ease data querying and analysis.

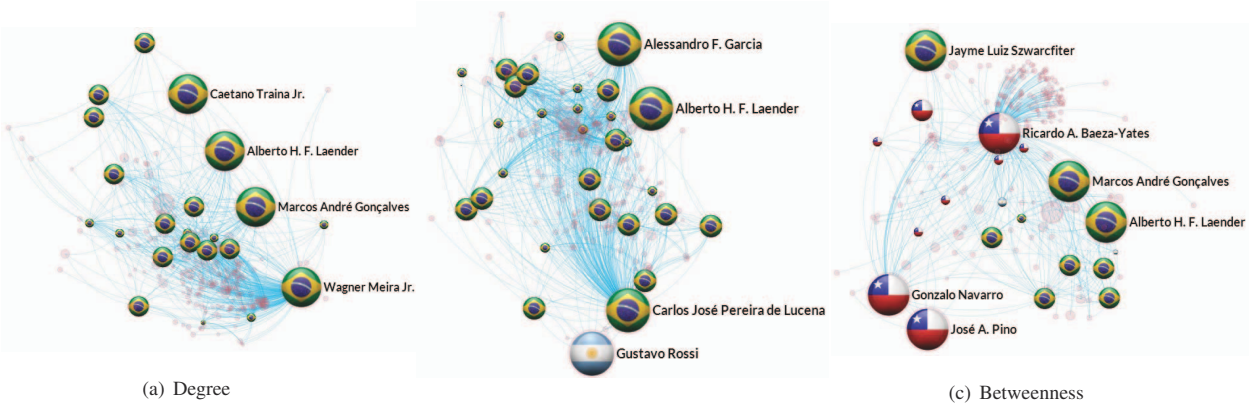
Basic Statistics. Here we present some basic statistics for LACompNet. The current graph is composed of 15601 vertices and 24722 edges. The total number of publications during the period between 1994 and 2013 is 18930 in 2887 venues, including books, journal articles and conference papers. Further, out of the 15601 vertices, 904 correspond to faculty members from the 35 Latin American institutions included in our study and the other 14697 are considered only as coauthors. We notice that 67.80% of them are from Brazilian, Chilean and Argentinian institutions.

Centrality Metrics. Graph centrality metrics are used to analyze the topological structure of a network, as well as to characterize the behavior and evolution pattern across time. In this work we use the *degree*, *closeness* and *betwenness centrality* metrics [11]. Tables I, II and III show, respectively, the list of the top-10 authors for each of these centrality metrics and Figures 1(a), 1(b) and 1(c) show the coauthorship networks of the top researchers according to each metric.

III. Network Analysis

In this section we analyze the coauthorship networks from each country as a whole and for each of the two periods of study (1994-2003 and 2004-2013). The general graph measures of LACompNet are presented in Table IV. Notice that the degree assortative coefficients are negative for all countries, which means that high-degree nodes tend

[4]http://www.saxproject.org/



(a) Degree

(b) Closeness

(c) Betweenness

Figure 1. Coauthorship networks of the top researchers according to each graph centrality metric.

Table I
TOP-10 AUTHORS BASED ON DEGREE CENTRALITY.

Author	Institutions	Degree
Wagner Meira Jr.	UFMG	0.0155
Ricardo A. Baeza-Yates	UCHILE	0.0151
Carlos José Pereira de Lucena	PUC-RIO	0.0133
Marcos André Gonçalves	UFMG	0.0130
Luigi Carro	UFRGS	0.0123
Alessandro F. Garcia	PUC-RIO	0.0115
Jano Moreira de Souza	UFRJ	0.0112
Silvio Romero de Lemos Meira	UFPE	0.0109
Jussara M. Almeida	UFMG	0.0107
Carlos A. Coello Coello	CINVESTAV	0.0106

Table II
TOP-10 AUTHORS BASED ON CLOSENESS CENTRALITY.

Author	Institutions	Closeness
Carlos José Pereira de Lucena	PUC-RIO	0.2453
Alberto H. F. Laender	UFMG	0.2425
Ricardo A. Baeza-Yates	UCHILE	0.2387
Nivio Ziviani	UFMG	0.2386
José Carlos Maldonado	ICMC-USP	0.2370
Simone Diniz Junqueira Barbosa	PUC-RIO	0.2339
Jussara M. Almeida	UFMG	0.2331
Wagner Meira Jr.	UFMG	0.2324
Marcos Andre Gonçalves	UFMG	0.2322
Thaís Vasconcelos Batista	UFRN	0.2312

Table III
TOP-10 AUTHORS BASED ON BETWEENNESS CENTRALITY.

Author	Institutions	Betweenness
Ricardo A. Baeza-Yates	UCHILE	0.0986
Carlos José Pereira de Lucena	PUC-RIO	0.0600
José Carlos Maldonado	ICMC-USP	0.0412
Jayme Luiz Szwarcfiter	UFRJ	0.0352
Wagner Meira Jr.	UFMG	0.0344
Antonio Alfredo Ferreira Loureiro	UFMG	0.0316
Carlos A. Coello Coello	CINVESTAV	0.0308
Jorge Urrutia	UNAM	0.0306
Sergio F. Ochoa	UCHILE	0.0295
Alberto H. F. Laender	UFMG	0.0289

Table IV
GRAPH MEASURES PER COUNTRY.

Country	Inst.	Graph Size $\|E\|$	Deg. Ass. Coeff.	Avg Clust.	Cliques [+]	LCC [*]
Argentina	4	1839	-0.371	0.247	1273	964
Brazil	8	17567	-0.485	0.286	11756	10401
Chile	5	2674	-0.306	0.195	1982	1755
Colombia	5	553	-0.448	0.105	479	252
Cuba	3	291	-0.310	0.217	189	54
Mexico	4	1729	-0.291	0.096	1515	1278
Peru	1	61	-0.597	0.047	57	22
Uruguay	1	379	-0.386	0.158	271	232
Venezuela	4	677	-0.329	0.207	505	289

[+] Cliques: Subsets of the vertex set $C \subseteq V$, such that for every two vertices in C, there exists an edge connecting the two.
[*] LCC: Largest Connected Component

to attach to low-degree nodes. In our case, this indicates that senior researchers frequently publish together with younger researchers and students. Observing the average clustering, we can confirm the trend regarding a larger number of collaborations in Brazil, contrasting to the fewest collaborations in Peru. Such differences w.r.t. critical mass in each country are confirmed by the number of cliques and the size of the largest connected components.

Table V shows an average relative increase of 57.50% w.r.t. the number of authors (including researchers and collaborators) and 69.11% w.r.t. the number of coauthorships. Colombia, Cuba, Peru and Uruguay are the countries that presented the largest increase in both metrics, since they are the countries for which we considered the smallest number of institutions and had the fewest researchers in the first period of analysis (1994-2003). On the other hand, Brazil presented the smallest relative increase, as a consequence of our study having considered only the top eight institutions in

Brazil (levels 6 and 7 according to CAPES[5], the Brazilian Ministry of Education agency in charge of graduate programs), which were already quite consolidated during both periods of study.

Table V
COAUTHORSHIP NETWORK GRAPH STATISTICS.

Country	1994 - 2003		2004 - 2013		Rel. Inc. %	
	Vertices	Edges	Vertices	Edges	Vertices	Edges
Argentina	1076	1450	1292	1830	20.07	26.20
Brazil	10113	16277	10560	17552	4.42	7.83
Chile	1661	2095	1958	2667	17.88	27.30
Colombia	248	259	505	545	103.62	110.42
Cuba	94	100	209	284	122.34	184
Mexico	1322	1455	1537	1724	16.26	18.48
Peru	27	25	65	61	140.74	144
Uruguay	182	219	290	379	59.34	73.05
Venezuela	408	507	542	663	32.84	30.76
Average	1681.22	2487.44	1884.22	2856.11	57.50	69.11

Research Networks. In order to assess the level of collaboration among the researchers in each network, here we present an analysis of the research networks that arose from the coauthorships and discuss how they evolved across time. We define a research network as a group of authors who have published together at least five papers in a decade [12]. We determined such networks by mining maximal sets [11] of authors. Table VI shows both the number of groups and their average size per country and decade. As we can see, all countries presented an increase w.r.t both indicators. The increase in the number of groups demonstrate that there is an increasing research density in the region, while the increase in the average size of the groups shows that researchers are cooperating more and there is a growing critical mass in the area. It is also interesting to notice that such trend appears for almost all countries regardless the number of papers published and number of research groups. Such evaluation also shows the increase of the critical mass in Latin America in terms of active countries, as demonstrated by the appearance of research groups in Colombia, as well as significant increases w.r.t. the number of groups in other countries. The only exception is Peru, which experienced a ten fold increase w.r.t. papers published after 2003, but there was no research group that published more than four papers. In fact, just one research group published four papers, four groups published two papers, and the other groups published just one paper each.

International Collaboration. Here we analyze the cooperations between Latin American countries as materialized by papers that contain authors from more than one country. We start by presenting, in Table VII, the number of papers published by authors from institutions located in a given country in the two periods of analysis. All countries experienced increases in the number of papers published. In Brazil and Argentina the increase was more than 280% in each country, in Mexico 181%, and in Chile 423%. Several countries experienced quite impressive increases: Colombia (1944%), Cuba (514%), Peru (933%), and Uruguay (510%). However, it is remarkable that the number of papers produced by these countries till 2003 was very small, being at most 29 papers in 10 years. Finally, Venezuela showed the worst increase (148%), maybe as a consequence of recent political changes there.

[5]http://www.avaliacaotrienal2013.capes.gov.br/

Table VI
GROUP ANALYSIS BY COUNTRY.

Country	1994 - 2003		2004 - 2013		Total
	Groups	Avg Size	Groups	Avg Size	
Argentina	24	2.17	108	2.72	132
Brasil	162	2.33	1060	2.60	1262
Chile	15	2.47	158	2.49	173
Colombia	0	0.00	23	2.43	23
Cuba	1	3.00	7	2.71	8
Mexico	25	2.48	96	2.54	121
Uruguay	1	2.00	16	2.50	17
Venezuela	6	2.00	24	2.38	30

Table VII
DISTRIBUTION OF PAPERS BY COUNTRY.

Country	1994 - 2003	2004 - 2013	Total
Argentina	307	1207	1514
Brazil	2654	10157	12811
Chile	396	2069	2465
Colombia	16	327	343
Cuba	21	129	150
Mexico	456	1280	1736
Peru	3	31	34
Uruguay	29	177	206
Venezuela	155	384	539

We then proceed and check the number of papers that were written by authors from more than one country, which is shown in Table VIII. A very first observation is that all cooperative work involved just two countries. We could not find in our dataset papers from authors who work on more than two countries. We also present, for each decade, the relative importance of these papers considering the total number of papers published by the countries being considered. Given that the total number of papers published by country A is p_A, country B is p_B, and both is p_{AB}, the relative importance of p_{AB} is given by $p_{AB}/(p_A + p_B - pAB)$.

We can observe that, before 2004, there were just two significant cooperations in Latin America, namely Argentina-Brazil and Brazil-Chile. We found two other cases, but they were not significative, although the cooperation Brazil-Peru was responsible for two of the three publications from Peru in that period. When we look at the last ten years, we can see that several new cooperations arose, but none of them was as significative as those from the previous decade, being most of them in the single digits in terms of number of papers. It is also remarkable that we had a

Table VIII
DISTRIBUTION OF JOINT PAPERS BY COUNTRY GROUPS.

Country	1994 - 2003		2004 - 2013		Total
	Papers	% Total	Papers	% Total	
Argentina, Brazil	42	1.49	47	0.44	89
Argentina, Chile	0	0.00	10	0.32	10
Argentina, Colombia	0	0.00	1	0.07	1
Argentina, Mexico	0	0.00	1	0.04	1
Brazil, Chile	19	0.65	26	0.23	45
Brazil, Colombia	0	0.00	7	0.07	7
Brazil, Peru	2	0.08	1	0.01	3
Brazil, Uruguay	0	0.00	4	0.04	4
Brazil, Venezuela	2	0.07	3	0.03	5
Chile, Colombia	0	0.00	1	0.04	1
Chile, Mexico	0	0.00	3	0.09	3
Colombia, Venezuela	0	0.00	1	0.15	1
Cuba, Mexico	0	0.00	5	0.36	5

reduction in the relative importance of the two cooperations from the previous decade, although the absolute number of papers increased (36% between Brazil and Chile and 11% between Argentina and Brazil). It is also worth noting new cooperative work between Argentina and Chile.

IV. CONCLUSIONS

In this paper we studied the coauthorship networks of CS research groups from 35 Latin American institutions. The study is based on data from DBLP and spans the period from 1994 to 2013. Our analysis shows that there has been a significant increase in the number of publications in the last decade, in particular when we consider countries such as Colombia, Uruguay and Venezuela, with less research tradition in the area. We also observed a consolidation of research groups in other countries such as Argentina, Brazil, Chile and Mexico.

We also identified the influential authors in the area, according to the centrality metrics considered, which show a predominance of Brazilian, Chilean and Mexican researchers that are in traditional research centers and were able to establish research groups. These findings may be useful for strenghtening the existing networks and fostering new collaborations with the researchers located in priviledged network locations. We also analyzed research networks that emerged from the coauthorships (i.e., groups of authors who published consistently together) and found that the number and size of such groups increased in almost all countries, showing a clear densification process. Regarding international collaboration, we found that there is still a lot to improve, since most of the collaborations happen between just Brazil-Chile and Argentina-Brazil, although there is some growth in the diversity of the collaborations.

As future work, we intend to study in greater detail how the networks are formed and to investigate the impact of the research topics on the evolution of CS research in Latin America.

ACKNOWLEDGMENTS

This work is partially funded by InWeb (MCT/CNPq/FAPEMIG grant 573871/2008-6), and by the authors' individual grants from CAPES, CNPq and FAPEMIG.

REFERENCES

[1] M. E. Newman, "The structure of scientific collaboration networks," *PNAS*, vol. 98, no. 2, p. 404, 2001.

[2] X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel, "Co-authorship networks in the digital library research community," *Information Processing & Management*, vol. 41, no. 6, pp. 1462–1480, 2005.

[3] J. Huang, Z. Zhuang, J. Li, and C. L. Giles, "Collaboration over Time: Characterizing and Modeling Network Evolution," in *Proc, of WSDM*, Stanford, CA, USA, 2008, pp. 107–116.

[4] G. Maia, P. O. S. V. de Melo, D. L. Guidoni, F. S. H. Souza, T. H. Silva, J. M. Almeida, and A. A. F. Loureiro, "On the analysis of the collaboration network of the Brazilian symposium on computer networks and distributed systems - 30 Editions of history," *J. Braz. Comp. Soc.*, vol. 19, no. 3, pp. 361–382, 2013.

[5] M. A. Nascimento, J. Sander, and J. Pound, "Analysis of SIGMOD's Co-authorship Graph," *SIGMOD Record*, vol. 32, no. 3, pp. 8–10, 2003.

[6] A. L. Barábasi, Z. Néda, E. Ravasz, A. Schubert, and V. T, "Evolution of the social network of scientific collaborations," *Physica A: Statistical Mechanics and its Applications*, vol. 311, no. 3-4, pp. 590–614, 2002.

[7] G. V. Menezes, N. Ziviani, A. H. F. Laender, and V. A. F. Almeida, "A Geographical Analysis of Knowledge Production in Computer Science," in *Proc. of WWW*, Madrid, Spain, 2009, pp. 1041–1050.

[8] V. P. Freire and D. R. Figueiredo, "Ranking in collaboration networks using a group based metric," *J. Braz. Comp. Soc.*, vol. 17, no. 4, pp. 255–266, 2011.

[9] J. P. Mena-Chalco, L. A. Digiampietri, F. M. Lopes, and R. M. C. Junior, "Brazilian Bibliometric Coauthorship Networks," *JASIST*, vol. 66, no. 7, 2014.

[10] M. Ley, "DBLP: Some Lessons Learned," *Proc. of VLDB*, vol. 2, no. 2, pp. 1493–1500, 2009.

[11] M. J. Zaki and W. Meira Jr, *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.

[12] J. F. Delgado-Garcia, A. H. F. Laender, and W. Meira Jr., "A Preliminary Analysis of the Scientific Production of Latin American Computer Science Research Groups," in *Proc. of AMW*, Cartagena de Indias, Colombia, 2014.