UMA ANÁLISE DA PRODUÇÃO CIENTÍFICA EM CIÊNCIA DA COMPUTAÇÃO NA AMÉRICA LATINA

JUÁN FELIPE DELGADO GARCIA

UMA ANÁLISE DA PRODUÇÃO CIENTÍFICA EM CIÊNCIA DA COMPUTAÇÃO NA AMÉRICA LATINA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Alberto Henrique Frade Laender
Coorientador: Wagner Meira Júnior

Belo Horizonte
Março de 2015

Delgado Garcia, Juán Felipe

D352a Uma análise da produção científica em Ciência da Computação na América Latina / Juán Felipe Delgado Garcia. — Belo Horizonte, 2015

xx, 76 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais- Departamento de Ciência da Computação.

Orientador: Alberto Henrique Frade Laender
Coorientador: Wagner Meira Júnior

1. Computação - Teses. 2. Redes de colaboração científica. 3. Mineração de dados (Computação) - Teses. 4. Bancos de dados - Teses. 5. Bibliometria - Teses. I. Orientador. II. Coorientador. III. Título.

CDU 519.6*72(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Uma Análise da Produção Científica em Ciência da Computação na América Latina

JUAN FELIPE DELGADO GARCIA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ALBERTO HENRIQUE FRADE LAENDER - Orientador
Departamento de Ciência da Computação - UFMG

PROF. WAGNER MEIRA JÚNIOR
Departamento de Ciência da Computação - UFMG

PROF. BERTHIER RIBEIRO DE ARAÚJO NETO
Departamento de Ciência da Computação - UFMG

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO
Departamento de Ciência da Computação - UFMG

PROF. ROBERTO MARCONDES CÉSAR JÚNIOR
Departamento de Ciência da Computação - USP

Belo Horizonte, 23 de março de 2015.

A Deus.

Aos meus pais, Maria Elizabeth Garcia C. e Pablo Emilio Delgado T.

À minha irmã Lina Maria Delgado G.

Agradecimentos

Os meus mais sinceros agradecimentos a todos que me apoiaram durante estes dois anos de amplas vivências e experiências que enriqueceram minha vida pessoal, profissional e acadêmica. Gostaria de agradecer especialmente

- A Deus que iluminou dia-a-dia o meu caminho e me abençoou com sua companhia;
- Aos meus pais que sempre quiseram o melhor para mim;
- À minha irmã por sua ajuda, comentários e revisões incondicionais e oportunas;
- Aos meus tios e primos que me apoiaram nesta decisão de vida;
- Ao meu tio Norberto Garcia-Cairasco, exemplo de vida;
- Ao Prof. Alberto H.F. Laender que me deu a oportunidade de mostrar meu trabalho, confiando sempre em mim e me apoiando para seguir em frente;
- Ao Prof. Wagner Meira Jr. que me abriu as portas do Brasil e da UFMG para começar uma nova etapa na minha vida;
- Ao DCC/ICEx que me deu as boas-vindas e todo o seu apoio para que minha estadia aqui fosse a melhor possível;
- Ao LBD como um todo, professores e estudantes, que me receberam com os braços abertos;
- Aos meus amigos da vida toda Hernán, Francisco, Carlos e Ana;
- Ao Diego, Ricardo, Ana, Edgar e Luz, colegas do cafezinho da Praça, debates e brincadeiras, pelos bons momentos de descontração
- À CAPES e InWeb, por meio do CNPq e FAPEMIG, por financiar parcialmente este trabalho.

Resumo

Diferentes trabalhos têm sido realizados com a intenção de comparar a produção científica brasileira e de outros países da América Latina na área de Ciência da Computação (CC) com a de países da América do Norte, Europa e Ásia. Mesmo assim, pouco ainda se sabe sobre a produção científica dos grupos de pesquisa dos países latino-americanos na área. Nesta dissertação, discutimos o perfil de produção científica dos grupos mais relevantes dos seguintes países latino-americanos: Argentina, Brasil, Chile, Colômbia, Costa Rica, Cuba, México, Paraguai, Peru, Uruguai e Venezuela. Para isso, coletamos dados dos pesquisadores de 48 grupos desses países que participam de programas de Pós-graduação em CC, em um período de 20 anos, de 1994 até 2013. Os dados foram coletados da DBLP (*Digital Bibliography & Library Project*), a maior biblioteca digital da área de Ciência da Computação. Nosso trabalho inclui análises estatísticas da produção científica de cada país, além de análises das redes de colaboração baseadas em métricas e propriedades de redes complexas. Nossos resultados mostram um crescimento estável da produção científica na última década na Argentina, Brasil, Chile e México, e um crescimento expressivo em países com pouca tradição na área como Colômbia, Cuba, Costa Rica, Paraguai e Peru. Particularmente, mostramos que a colaboração científica entre grupos de prestígio na região produziu um aumento na média geral da taxa de publicação de cada grupo, além de gerar publicações de melhor qualidade. Além disso, comprovamos que as colaborações são um mecanismo importante para o aumento da taxa de publicações. Mostramos ainda que existe uma correlação entre as métricas centralidade de grau e centralidade de intermediação com o total de publicações dos pesquisadores. Igualmente, comprovamos que as redes de colaboração seguem a lei de potência, onde há evidência de poucos pesquisadores com um grau maior dentro da estrutura da rede que atuam ao mesmo tempo como "hubs" na rede.

Palavras-chave: Redes de Colaboração Científica, Mineração de Dados, Bancos de Dados, Bibliometria.

Abstract

Different studies have been conducted aiming to compare the scientific production in Computer Science (CS) of Brazil and other Latin American countries with that of different countries in North America, Europe and Asia. However, little is still known about the scientific production of researcher groups from these countries in the area. In this MSc dissertation, we study the profile of the most important research groups from Argentina, Brazil, Chile, Colombia, Costa Rica, Cuba, Mexico, Paraguay, Peru, Uruguay and Venezuela in a period of 20 years, from 1994 to 2013. To do so, our study addresses researchers of 48 graduate programs in CS from academic institutions in Latin American. The data was collected from DBLP (Digital Bibliography & Library Project), the largest CS digital library. Our work includes a statistical analysis of the scientific production of each country, as well as analyses of the coauthorship networks based on metrics and properties of complex networks. Our results show a stable growth of the scientific production in the last decade in Argentina, Brazil, Chile and Mexico, and a significant growth in countries with less tradition in CS such as Colombia, Cuba, Costa Rica, Paraguay and Peru. Particulary, we show that the scientific collaboration among prestigious research groups in the region has resulted in better quality publications and an increase in terms of publication rate of each research group. Thus, we confirm that collaborations are an important mechanism to increase the rate of publications. In addition, we show that exists a correlation between the degree and closeness centrality metrics with the total number of publications of the researchers. We also confirm that scientific collaboration networks follow the power law, where there is evidence of few researchers with a large degree within the network structure that at act as hubs on the network.

Keywords: Scientific Coauthorship Networks, Data Mining, Databases, Bibliometrics.

Lista de Figuras

Lista de Tabelas

Sumário

Lista de Abreviaturas

Abreviatura	*Descrição*
C_a	*Coeficiente de agrupamento*
C_{al_i}	*Coeficiente de agrupamento local*
C_g	*Tamanho do componente gigante*
l	*Comprimento do caminho médio*
d	*Diâmetro do componente gigante*
k	*Grau de um vértice*
$< k >$	*Grau médio*
Cp	*Centralidade de proximidade*
Ci	*Centralidade de intermediação*
I_T	*Índice T*
I_S	*Índice S*

Capítulo 1

Introdução

1.1 Motivação

Conforme dados do SCImago Journal & Country Rank (SJR[1]), o Brasil é responsável por $47,91\%$ da produção científica na América Latina, com uma contribuição de aproximadamente 529.841 artigos científicos no período $1996 - 2013$ em todas as áreas do conhecimento. Especificamente, na área de Ciência da Computação, o Brasil contribuiu com $48,77\%$ dos artigos científicos produzidos no mesmo período, abrangendo subáreas como Inteligência Artificial, Teoria da Computação, Computação Gráfica, Visão Computacional, Arquitetura e Hardware, Interação Humano-Computador, Sistemas de Informação, Processamento de Sinais e Engenharia de Software, segundo estatísticas geradas a partir da plataforma SCImago SJR[2]. A Tabela 1.1 mostra estatísticas sobre a produção científica em Ciência da Computação nos países da América Latina no período $1996 - 2013$.

Nos últimos anos, a produção científica na área da Ciência da Computação cresceu no mundo todo. Ainda considerando dados do SCImago SJR de 2004 a 2013, a Tabela 1.2 mostra o total de artigos científicos por regiões nesse período; de fato o crescimento em cada uma das regiões foi bastante expressivo: $11,25\%$ na América do Norte, $70,83\%$ na América Latina, $116,18\%$ na Europa Ocidental e $170,57\%$ em outras regiões. Embora esses números reflitam somente publicações que aparecem em periódicos, eles mostram que a área de Ciência de Computação é altamente produtiva com muitos grupos ativos espalhados pelo mundo.

[1] http://www.scimagojr.com/countryrank.php

[2] `http://www.scimagojr.com/countryrank.php?area=1700&category=0&region=Latin+America&year=all&order=it&min=0&min_type=it`

País	Ranking Mundial	# Artigos	# Citações	h-index
Brasil	16	35.778	138.660	106
México	32	16.664	62.817	81
Argentina	48	5.651	24.504	53
Chile	49	5.372	35.005	68
Colômbia	55	3.565	7.443	33
Venezuela	68	1.736	8.217	36
Cuba	71	1.277	6.239	37
Porto Rico	80	845	2.870	26
Uruguai	83	738	2.547	23
Peru	91	331	785	14
Costa Rica	102	216	654	12
Equador	105	206	1.189	14
Paraguai	123	79	303	8
Bolívia	132	41	193	8
El Salvador	138	30	82	6
Nicarágua	152	17	103	3
República Dominicana	159	13	76	4
Honduras	163	11	25	3

Tabela 1.1: Estatísticas sobre a produção científica na área de Ciência da Computação na América Latina no período 1996-2013.

Região	Países	2004	2013
América do Norte	Canada, USA	41.105	45.730
América Latina	Brasil, México, Argentina, Chile, Colômbia	2.793	6.038
Europa Ocidental	Alemanha, Franca, Itália, Espanha	28.245	48.251
Outras regiões	Austrália, China, Coreia, Índia, Polônia	24.239	65.584

Tabela 1.2: Crescimento da produção científica na área de Ciência da Computação em diferentes regiões durante os anos 2004 a 2013.

De acordo com as estatísticas apresentadas, a produção científica brasileira tem crescido nos últimos 25 anos em diferentes áreas do conhecimento como Farmacologia, Química, Neurociências, Bioquímica e Biologia Molecular e Psiquiatria, como mostrado por de Almeida & Guimarães [2013]. Esse resultado deve-se em grande parte a um esforço planejado do governo brasileiro por meio de agências de fomento como a Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES[3]) e

[3] `www.capes.gov.br/`

o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq[4]), além de outras agências estaduais que apoiam o financiamento de estudos de nível superior. Assim, os investimentos do governo foram fundamentais para o desenvolvimento da pesquisa científica no Brasil como reportado por Hermes-Lima et al. [2008]. Igualmente, o crescimento de sua economia como fator chave nos últimos 20 anos tem feito com que a sua produção científica corresponda a mais de dois terços de toda a produção da América do Sul, embora seja muito semelhante às da Argentina, Uruguai e Chile em termos de artigos per capita [Van Noorden, 2014]. Nesse contexto, o Brasil tem-se tornado um importante motor de desenvolvimento científico, além de tornar-se um exemplo na América Latina no avanço da ciência e tecnologia, com programas de Pós-graduação de alta qualidade acadêmica e padrão internacional. De fato, mais e mais grupos de pesquisa são criados a cada dia, em busca de desenvolver pesquisa científica que leve à geração de novos conhecimentos para toda a comunidade.

Embora muitos trabalhos tenham sido elaborados comparando a produção de programas e grupos de pesquisa brasileiros com a de universidades americanas, europeias e asiáticas [Wainer et al., 2009], os resultados mostraram um alto nível de desenvolvimento e qualidade dos programas e grupos de pesquisa brasileiros [Laender et al., 2008]. Do mesmo modo, na área de Ciência da Computação, Laender et al. [2008] mostram que os programas brasileiros possuem boa produtividade seguindo os mesmos padrões em termos de taxas de publicação dos programas internacionais de países da América do Norte e Europa. Da mesma forma, Mena-Chalco et al. [2014] apresentam uma análise da rede de colaboração brasileira dos pesquisadores que se encontram na plataforma Lattes do CNPq, visando caracterizar e analisar as redes de coautoria nas principais grandes áreas do conhecimento.

Contudo, não encontramos na literatura um estudo que forneça um retrato da produção científica na área de Ciência da Computação nos países da América Latina e compare a estrutura das redes de colaboração e a produção científica desses países e do Brasil como maior produtor de artigos científicos na região. É de vital importância conhecer como têm-se desenvolvido esses grupos ao longo do tempo e como é a estrutura de cada grupo de pesquisa como fator importante para identificar futuras colaborações entre instituições e fortalecer a área de Ciência da Computação na região.

Portanto, o foco central deste trabalho é fazer uma comparação dos grupos em

[4] `http://www.cnpq.br/`

Ciência da Computação da América Latina, considerando que outros estudos realizados já fazem uma comparação dos grupos brasileiros com os de outros países [Laender et al., 2008]. Levando em consideração esse cenário, nesta dissertação apresentamos uma análise da produção científica dos principais grupos de pesquisa da América Latina na área de Ciência da Computação. Nossa análise está baseada em dados de 20 anos coletados da DBLP[5] [Ley, 2009], a principal fonte de dados bibliográficos da área, compreendendo grupos de 48 instituições da Argentina, Brasil, Chile, Colômbia, Costa Rica, Cuba, México, Paraguai, Peru, Uruguai e Venezuela. Nesse contexto, o uso de ferramentas para análise de redes complexas [Albert & Barabási, 2002; Barabási et al., 2002; Boccaletti et al., 2006; Newman, 2001b] nos fornece um conjunto de métricas importantes para analisar a estrutura e a evolução das redes de colaboração desses países ao longo do tempo. Sendo assim, definimos uma rede de coautoria como um tipo especial de rede social em que os autores estão conectados entre si se têm pelo menos uma publicação em conjunto. Neste contexto, este trabalho fornece uma visão de como essas redes evoluíram ao longo do tempo, quais são os principais pesquisadores de cada um dos grupos e como eles estão posicionados na rede.

Finalmente, uma das principais contribuições desta dissertação é fazer uma caracterização que permita conhecer como as redes estão organizadas, além de identificar novos grupos-alvo que possam enriquecer a pesquisa que atualmente se faz no Brasil, com o estabelecimento de cooperações científicas e alianças estratégicas. Além disso, os resultados desta pesquisa podem fornecer informação vital para a busca de novos estudantes que fortaleçam o processo de internacionalização das universidades brasileiras.

1.2 Trabalhos Relacionados

Diferentes trabalhos têm sido realizados a fim de analisar o crescimento de áreas de pesquisa no Brasil. de Almeida & Guimarães [2013] identificam as áreas de maior crescimento que geram novos conhecimentos no Brasil, assim como a sua evolução e a relação entre a produção científica e o Programa Nacional de Pós-Graduação (PNPG[6]). Nesse estudo é feita uma comparação com os países de maior incremento na produção de artigos científicos no período compreendido entre 2006 e 2010. A comparação entre os períodos 1981-1985 e 2006-2010 mostra que o Brasil faz parte de um pequeno grupo de países (Coréia do Sul, China, Irã, Turquia, Taiwan, Cingapura,

[5] http://www.informatik.uni-trier.de/~ley/db

[6] http://www.capes.gov.br/sobre-a-capes/plano-nacional-de-pos-graduacao

Portugal, Hong Kong, Espanha, México e Grécia) que alcançaram altos índices de crescimento na produção científica nos últimos 30 anos, ou seja, que cresceram pelo menos quatro vezes a média mundial.

Mena-Chalco et al. [2014] apresentam um estudo abrangente da comunidade científica brasileira ao caracterizar e explorar suas principais redes de coautoria, com o intuito de obter uma compreensão em profundidade das estruturas da rede, bem como da dinâmica (comportamento social) entre os pesquisadores nas oito principais áreas de conhecimento brasileira: ciências agrárias, ciências biológicas, ciências exatas e da terra, ciências humanas, ciências sociais aplicadas, ciências da saúde, engenharia e linguística, letras e artes. Nesse trabalho, a representação de redes de coautoria mediante grafos e o uso de técnicas bibliométricas representam um valioso instrumento para proporcionar um melhor entendimento da dinâmica de colaborações entre pesquisadores e grupos. Do mesmo modo, Maia et al. [2013] apresentam uma detalhada análise da estrutura e evolução da rede de coautoria do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC) ao longo de suas 30 edições, identificando as regiões do Brasil que atraem mais autores, os pesquisadores com funções centrais na rede e as colaborações entre autores de diferentes regiões geográficas.

Digiampietri et al. [2014] apresentam uma caracterização de 37 programas brasileiros de graduação em Ciência da Computação usando dados coletados a partir da plataforma Lattes, que compreende 732 professores e 17.976 publicações em conferências e periódicos. Para a análise das características das redes acadêmicas foram usadas diversas métricas que mostram que os programas com maior presença na topologia da rede de coautoria também apresentam maior produtividade em pesquisa no período de estudo (2004 - 2009). Da mesma forma, Coutinho et al. [2012] analisam a produção científica do Brasil na área de educação. O Brasil, como país em desenvolvimento, tem atravessado um processo de crescimento em diversas áreas, incluindo o desenvolvimento da pesquisa científica, especialmente quando comparado com países desenvolvidos da Europa e América do Norte com um incremento do número de pesquisadores, programas de Pós-graduação e grupos de pesquisa na última década.

Nessa mesma linha, Wainer et al. [2009] apresentam um estudo comparativo da produção científica do Brasil em Ciência da Computação com outros países no período 2001 a 2005. Os países envolvidos na comparação incluem alguns da América

Latina (Argentina, Chile e México), do grupo BRIC (Rússia, Índia e China) e outros como Coréia do Sul, Austrália e USA. Os resultados mostraram que o Brasil é de longe o maior produtor de artigos científicos na América Latina, ficando próximo a países da Comunidade Europeia como Espanha e Itália, e muito perto da Índia e Rússia.

O estudo de Laender et al. [2008] visou analisar a qualidade dos programas de Pós-graduação em Ciência da Computação do Brasil, mostrando que esses programas se comparam bem em termos de taxa de publicação e número de graduados com programas da América do Norte e Europa. A importância da bibliometria nesse estudo é evidente, pois permite determinar o nível de internacionalização desses programas com relação à sua produção científica. O estudo mostrou que os principais programas de Pós-graduação em Ciência de Computação do Brasil seguem um bom padrão de produção quando comparados aos programas europeus e norte americanos escolhidos para o estudo. Foi mostrado também que os programas brasileiros seguem índices de publicação internacionais de mais de dois artigos em conferência por artigo em periódico. Portanto, qualquer processo de avaliação não pode ignorar as publicações em conferências uma vez que estas são veículos fundamentais para a divulgação da pesquisa em Ciência da Computação.

Como parte dos esforços para fornecer ferramentas que ajudem a visualizar a produção acadêmica brasileira, o projeto *CiênciaBrasil*[7], conduzido no Laboratório de Bancos de Dados[8] do Departamento de Ciência de Computação da UFMG, produziu uma serie de ferramentas que permitem visualizar a rede de coautoria brasileira com base em dados da Plataforma Lattes a fim de comparar diferentes grupos de pesquisa dentro do programa dos Institutos Nacionais de Ciência e Tecnologia (INCT)[9] [Laender et al., 2011a,b]. Do mesmo modo, Kurosawa & Takama [2012] apresentam um sistema de visualização de uma rede bibliográfica que é composta por uma rede de coautoria, uma rede de citações e uma rede de co-citações, fornecendo informações detalhadas sobre autores específicos, artigos e conferências que são úteis para diferentes estudos. Da mesma forma, Huang & Huang [2006] propõem uma nova abordagem para a coleta, análise e visualização de dados de coautoria. Esse trabalho foi avaliado a partir de dados da DBLP e permite compreender como foi a colaboração acadêmica dos pesquisadores em um período passado a traves de suas co-publicações.

[7] `http://pbct.inweb.org.br/pbct/`
[8] `http://www.lbd.dcc.ufmg.br/`
[9] `http://inct.cnpq.br/`

Finalmente, as redes de colaboração têm sido estudadas amplamente ao longo dos anos [Huang et al., 2008; Liu et al., 2005; Maia et al., 2013; Nascimento et al., 2003; Newman, 2001b], fornecendo uma interessante visão das comunidades acadêmicas por detrás delas. Entre os trabalhos pioneiros, Newman [2001b] analisa três comunidades científicas (Ciência da Computação, Física e Biomedicina) e apresenta várias características estruturais e topológicas delas, através do mapeamento das publicações em periódicos importantes de Matemática e Neurociências em um período de oito anos (1991-1998). Barabási et al. [2002] inferem a dinâmica e os mecanismos estruturais que governam a evolução e topologia das redes de coautoria de duas comunidades (Matemática e Neurociências) com base em várias métricas de redes complexas. Do mesmo modo, Menezes et al. [2009] estudaram a produção em Ciência da Computação usando redes de coautoria de 30 programas de Pós-graduação em diferentes regiões do mundo. O conjunto de dados consistiu em 176.537 autores e 352.766 publicações distribuídas em 2.176 veículos (conferências e periódicos), enquanto que diferentes métricas de redes complexas foram aplicadas para entender o comportamento de cada uma das redes. Os resultados mostraram que a produção do conhecimento mudou diferentemente para cada região, e que o uso de redes complexas pode ser uma alternativa eficaz para a compreensão e análise do processo de produção do conhecimento em Ciência da Computação em diferentes regiões geográficas.

1.3 Contribuições

As principais contribuições desta dissertação são:

- Desenvolvimento de uma plataforma computacional denominada LACompNet[10] (Latin American Computer Science Network) para coleta, tratamento e visualização dos dados dos principais grupos de pesquisa em Ciência da Computação da América Latina;

- Caracterização da produção científica em Ciência da Computação dos principais países latino-americanos, cujos resultados preliminares foram apresentados no 8^{th} Alberto Mendelzon Workshop on Foundations of Data Management [Delgado-Garcia et al., 2014a];

- Análise detalhada das redes de coautoria em Ciência da Computação de 31 instituições acadêmicas latino-americanas, cujos resultados preliminares foram apresentados no 9^{th} Latin American Web Congress [Delgado-Garcia et al., 2014b] e

[10] `http://tortuga.lbd.dcc.ufmg.br/LACompNet`

também na sessão especial de posteres do projeto InWeb realizada no mesmo evento.

1.4 Organização da Dissertação

Os demais capítulos desta dissertação estão organizados como se segue. O Capítulo 2 apresenta o referencial teórico de redes complexas e as principais métricas usadas ao longo da dissertação. O Capítulo 3 descreve a plataforma que suporta o tratamento dos dados e a análise das redes de coautoria por países e por instituições. O Capítulo 4 descreve em detalhe cada uma das redes de coautoria dos países envolvidos no estudo, fazendo uma caraterização dessas redes e uma análise de cada grupo envolvido. O Capítulo 5 apresenta várias visualizações das redes de coautoria usando a plataforma desenvolvida. Finalmente, o Capítulo 6 apresenta conclusões e diretrizes para trabalhos futuros.

Capítulo 2

Redes Complexas

Este capítulo apresenta os fundamentos teóricos de redes complexas necessários para analisar as redes de coautoria de cada um dos países abordados nesta dissertação.

2.1 Introdução

Atualmente, o conceito de redes complexas está presente na maioria das estruturas do mundo real, além de estabelecer uma relação com diferentes campos da ciência. A importância desse tipo de estrutura deve-se ao fato que permite descrever o comportamento de diferentes sistemas e seus relacionamentos através de métricas que definem propriedades estruturais dessas redes. Como exemplos de redes complexas, temos: a Internet, a World Wide Web, redes sociais, redes de organizações e de relações de negócios entre empresas, redes neurais, redes metabólicas, redes de distribuição, redes de citações, e muitas outras.

Um caso particular desse tipo de rede são as redes sociais. Segundo Easley & Kleinberg [2010], as redes sociais são definidas pelas interações entre amigos ou grupos de pessoas. Essas redes têm crescido em complexidade ao longo da história, devido aos avanços tecnológicos que facilitam a comunicação global e interação digital. Do mesmo modo, Newman [2001b] define uma rede social como uma coleção de indivíduos ou grupos de indivíduos conectados por um tipo de relação existente entre eles. Neste caso, os indivíduos ou grupos de indivíduos são chamados de *atores* e as relações entre eles são chamadas de *laços*. Esse tipo de rede pode ser representado por um grafo.

A Figura 2.1 ilustra uma rede composta por sete atores e seis laços ou relações entre eles.

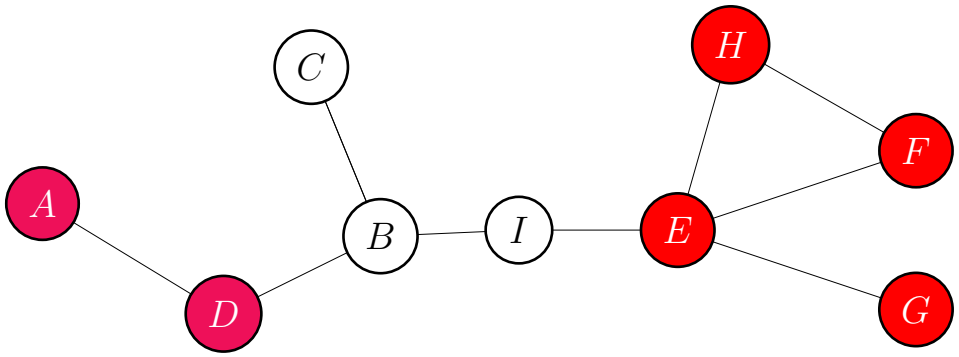
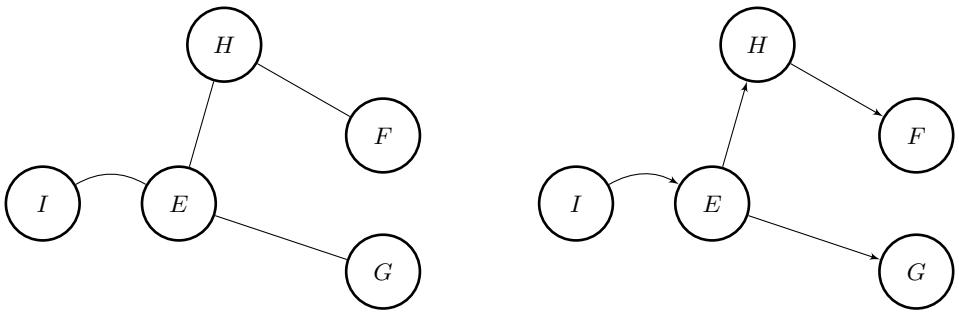


Figura 2.1: Estrutura da rede $G = (V, E)$.

Formalizando, esse tipo de rede é representado topologicamente como um grafo $G = (V, E)$ tal que, V é um conjunto finito de vértices $v_1, v_2, v_3, ..., v_n$ e E o conjunto de arestas $e_1, e_2, e_3, ..., e_n$ que enlaçam os pares (v_i, v_j). Neste contexto, se dois pares de vértices (v_i, v_j) não ordenados estão enlaçados como ilustrado na Figura 2.2a, se diz que o grafo é não dirigido; em contraste se dois pares de vértices ordenados estão enlaçados como ilustrado na Figura 2.2b se diz que o grafo é dirigido [Zaki & Meira Jr, 2014].



(a) Não Dirigido (b) Dirigido

Figura 2.2: Exemplo de grafo não dirigido e dirigido.

O número de nós de um grafo G é dado por $|V| = n$ e é denominado de ordem do grafo. O número de arestas de um grafo é dado por $|E| = m$ e determina o tamanho do grafo [Zaki & Meira Jr, 2014].

Do mesmo modo, uma rede de coautoria é um tipo especial de rede social em que os atores representam autores e os laços indicam que os autores têm pelo menos uma coautoria em uma publicação em conjunto. Devido à grande quantidade de dados bibliográficos disponibilizados hoje na Web, redes de coautoria têm sido amplamente

estudadas ao longo dos últimos anos [Bazzan & Argenta, 2011; Delgado-Garcia et al., 2014a; Digiampietri et al., 2014; Franceschet, 2011; Huang et al., 2008; Liu et al., 2005; Maia et al., 2013; Nascimento et al., 2003; Newman, 2001b; Rodriguez & Pepe, 2008], proporcionando uma visão interessante das comunidades acadêmicas por trás delas.

2.2 Métricas de Centralidade

As métricas de centralidade são usadas para determinar a importância dos vértices ou arestas de um determinado grafo. Nesta seção apresentamos três métricas de centralidade: centralidade de grau, centralidade de proximidade (*closeness*) e centralidade de intermediação (*betweenness*) [Wasserman, 1994] ilustradas na Figura 2.3.

2.2.1 Centralidade de Grau (*Degree*)

O grau k de um vértice é uma das métricas mas simples para medir o grau de importância de um vértice em uma rede, e consiste do número de arestas incidentes em um vértice no caso de grafos não dirigidos, como os estudados nesta dissertação. Para grafos dirigidos, pode-se contar o número de arestas entrantes *(Indegree)* e o número de arestas saídas do mesmo vértice *(Outdegree)*.

2.2.2 Centralidade de Proximidade (*Closeness*)

A métrica centralidade de proximidade define quão perto se encontra um vértice em relação aos demais vértices do grafo [Zaki & Meira Jr, 2014], ou seja, o valor dessa métrica para um vértice v_i expressa a proximidade desse vértice em relação aos demais, sendo calculado pela soma de todas as distâncias do vértice v_i até v_j de acordo a sua proximidade. Neste caso, quanto menor a distância maior a proximidade. Esta métrica é definida pela Equação (2.1),

$$Cp_{(v_i)} = \frac{1}{\sum_{j \in v} d(v_i, v_j)} \tag{2.1}$$

Em redes de colaboração ou coautoria, conhecer quais são os vértices com maior proximidade possibilita conhecer quais são os autores que podem disseminar novo conhecimento através da rede com maior rapidez.

2.2.3 Centralidade de Intermediação (*Betweenness*)

A métrica centralidade de intermediação quantifica o número de vezes que um vértice atua como ponte ou intermediário ao longo do caminho mais curto entre outros vértices [Sun & Tang, 2011]. Esta métrica é definida pela Equação (2.2),

$$Ci_{(v_i)} = \sum_{s,t \in v} \frac{\sigma(s,t|v_i)}{\sigma(s,t)} \tag{2.2}$$

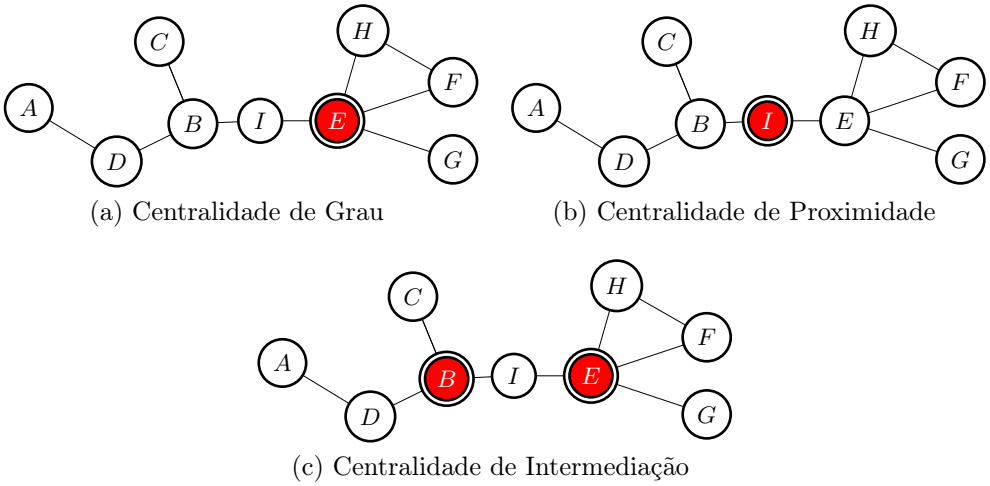
onde $\sigma(s,t)$ é o número total de caminhos mínimos desde s até o vértice t e $\sigma(s,t|v_i)$ é o total desses caminhos mínimos que passam pelo vértice v_i. No caso desta métrica em redes de coautoria, ela nos permite conhecer aqueles vértices da rede que têm o papel de pontes de transferência de nova informação entre diferentes grupos de pesquisa. Está métrica pode ser estendida também às arestas. Neste caso a métrica de intermediação de uma aresta é definido como o número de caminhos mais curtos entre pares de vértices que passam através dessa aresta [Newman & Girvan, 2004]. Como exemplo de uso dessa métrica, nesse mesmo estudo de Newman e Girvan, a detecção de comunidades científicas em redes de colaboração cientifica é feita através da remoção das arestas de maior grau de intermediação, gerando a criação de pequenos componentes fortemente conectados.

2.2.4 Exemplos das Métricas de Centralidade

Considerando o grafo da Figura 2.1, podemos identificar os vértices que se encontram mas bem posicionados de acordo com cada uma das três métricas de centralidade previamente definidas. Na Figura 2.3a, o vértice E é o de maior grau no grafo, ou seja, possui o maior número de arestas incidentes nele. A Figura 2.3b mostra o vértice de maior proximidade com os seus vizinhos e, por último, a Figura 2.3c mostra os dois vértices com os maiores valores de intermediação nos quais o fluxo de informação passa constantemente por eles; em outras palavras, se removermos esses vértices do grafo encontraremos duas comunidades ou grupos de pesquisa possivelmente de áreas diferentes no caso de uma rede de colaboração científica.

2.3 Propriedades das Redes

Para analisar redes complexas, faz-se necessário entender propriedades importantes que as caracterizam e como elas estão estruturadas em termos dos seus principais componentes. Essas propriedades possuem um papel chave ao se comparar diferentes



(a) Centralidade de Grau (b) Centralidade de Proximidade

(c) Centralidade de Intermediação

Figura 2.3: Métricas de centralidade no grafo G.

redes ou entender a semelhança entre elas. Entre as propriedades mais importantes temos: distribuição dos graus, coeficiente de agrupamento e assortatividade.

2.3.1 Distribuição dos Graus

Conforme já definido, o grau de um vértice corresponde ao número de arestas que incidem sobre ele, embora nem todos os vértices de uma rede têm o mesmo número de arestas ou mesmo grau. A propagação dos graus é caracterizada por uma função de distribuição $P(k)$ que expressa a probabilidade de que ao se escolher um vértice aleatoriamente, este tenha exatamente k arestas incidentes nele [Albert & Barabási, 2002; Newman, 2001b].

Particularmente, em um grande número de redes, tais como a World Wide Web [Albert et al., 1999], a Internet [Faloutsos et al., 1999], as redes metabólicas [Jeong et al., 2000], e as redes de contato sexual [Jones & Handcock, 2003; Liljeros et al., 2001], a distribuição do grau dos seus vértices tem uma cauda pesada seguindo a lei de potência $P(k) \sim k^{-\gamma}$ como também ocorre com as redes de colaboração científica [Albert & Barabási, 2002; Barabási et al., 2002; Newman, 2001b]. De fato, uma métrica comumente utilizada para comparar diferentes redes é o expoente $-\gamma$ obtido geralmente por meio de uma regressão linear. Valores típicos para aquele exponente ficam entre 1,0 e 3,5 [Benevenuto et al., 2011]. Para verificar a acurácia da regressão, é comum medir o coeficiente de determinação R^2 [McCool, 2003] que varia entre 0 e 1. Por fim, quanto mais próximo o valor de R^2 for de 1 (regressão perfeita), menor será a diferença entre o modelo de regressão e os dados reais [Benevenuto et al., 2011].

2.3.2 Grau Médio

O grau médio $< k >$ de um grafo G é a média dos graus de cada vértice, conforme expresso pela Equação 2.3,

$$< k >= \frac{1}{N} \sum_{i \in v} k_i \tag{2.3}$$

onde v é o conjunto de vértices do grafo G e k_i é o grau do vértice i.

2.3.3 Coeficiente de Agrupamento

Por definição, o coeficiente de agrupamento também conhecido como de *transitividade*, é uma propriedade típica das redes de amizade, onde dois indivíduos com amigos em comum são susceptíveis de se conhecerem [Wasserman, 1994]. Em outras palavras, se o vértice A está conectado ao vértice B e o vértice B está conectado ao vértice C, então existe uma probabilidade elevada de que o vértice A também esteja conectado ao vértice C [Newman, 2003b]. Por conseguinte, o coeficiente de agrupamento de um vértice v_i é uma medida de densidade do número de arestas entre os vizinhos de v_i [Zaki & Meira Jr, 2014], de tal forma que o coeficiente de agrupamento local pode ser quantificado pela Equação 2.4,

$$C_{al_i} = \frac{2e_i}{n_i(n_i - 1)} \tag{2.4}$$

onde e_i é o número de arestas entre os vizinhos de i e n_i é o número de vizinhos do vértice i. Assim, o coeficiente de agrupamento global é definido por meio da Equação 2.5,

$$C_a = \frac{1}{N} \sum_{i \in v} C_{al_i} \tag{2.5}$$

onde C_a é a média de todos os coeficientes de agrupamento local de cada vértice no grafo.

2.3.4 Assortatividade

Esta propriedade define a preferência dos vértices de um grafo G de se conectarem com outros vértices que são similares em termos do seu grau [Newman, 2003a]. O coeficiente por sua vez pode variar entre -1 e 1. Assim, quando os vértices de maior grau tendem a se conectar com vértices similares a ele, se diz que o grafo tem assortatividade *positiva*; no entanto, quando vértices de maior grau tendem a se conectar com vértices de grau menor, se diz que o grafo tem assortatividade *negativa*.

2.3.5 Componente Gigante

Um componente de um grafo é um subgrafo com a característica de que existe um caminho entre um vértice e qualquer outro vértice desse subgrafo [Liu et al., 2005]. Em outras palavras, um componente permite determinar o número de vértices que se encontram conetados na rede toda. De fato, o maior componente conectado se denomina componente gigante C_{gg}, que cobre a maior parte da rede. A Figura 2.4 ilustra o maior componente conectado de uma rede de coautoria. Esse componente inclui 1.968 vértices abrangendo 94.25% da rede. Comumente, uma rede de coautoria consiste de muitos componentes desconectados, onde grupos de pesquisadores se juntam mas não estão sempre ligados à rede inteira.

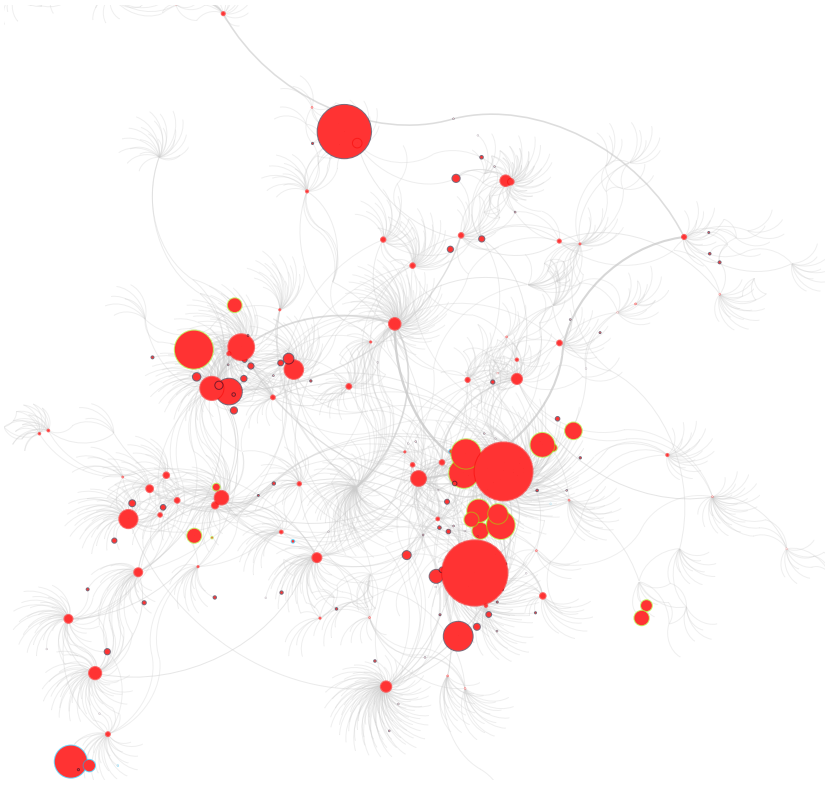


Figura 2.4: Maior componente conectado da rede de coautoria das instituições chilenas.

2.3.6 Diâmetro

Segundo Newman [2003b], o diâmetro d_g de uma rede é o comprimento em número de arestas do caminho geodésico mais longo entre quaisquer pares de vértices. Em redes do mundo real, o diâmetro de um grafo desconectado é infinito, embora ele possa ser definido como a maior distância de cada um do seus componentes conetados [Albert & Barabási, 2002].

2.3.7 Comprimento do Caminho Médio

Comprimento do caminho médio é um conceito em topologia de rede que é definido como o número médio de passos ao longo dos caminhos mais curtos para todos os possíveis pares de nós da rede. É uma medida da eficiência de informação ou de transporte sobre uma rede. Geralmente, o comprimento do caminho médio da rede toda é determinado pelo componente gigante.

2.4 Modelos ou Tipos de Redes

Diferentes modelos de rede têm sido estudados amplamente na literatura, entre eles os modelos de Erdös & Rényi [1959, 1960], de Watts & Strogatz [1998], de Barabási & Albert [1999] e o modelo de Barabási et al. [1999], também conhecido como redes livres de escala.

Nesta dissertação iremos concentrar-nos em dois modelos que estão presentes na maioria das redes de colaboração científica e que se encaixam perfeitamente no tema central deste estudo, a saber: redes livres de escala (Subseção 2.4.1) e redes de mundo pequeno (Subseção 2.4.2).

2.4.1 Redes Livres de Escala

Uma rede é dita livre de escala se sua distribuição de grau é livre de escala, ou seja, se sua distribuição de grau segue uma lei de potência [Barabási et al., 1999]. Consequentemente, em uma rede livre de escala, os graus dos vértices não são nada parecidos uns com os outros, pois podemos ter vértices com graus muito maiores do que a média com probabilidade não desprezível. De fato, muitas redes reais possuem esta propriedade e, por isto, são chamadas de redes livres de escala [Figueiredo, 2011].

2.4.2 Redes de Mundo Pequeno (*Small World*)

O modelo de redes de mundo pequeno foi descrito pela primeira vez pelo escritor de origem húngara Frigyes Karinthy em 1929 e provado experimentalmente por Milgram [1967]. A ideia principal deste modelo é que duas pessoas arbitrárias estão conectadas por apenas seis graus de separação, ou seja, o diâmetro do grafo correspondente não é muito maior do que seis. Em 1998, Watts e Strogatz [Watts & Strogatz, 1998; Watts, 1999a,b] publicaram o primeiro modelo de rede de mundo pequeno. Nesse modelo de

Watts e Strogatz, foi demostrado que com a adição de apenas um pequeno número de enlaces de longo alcance em um grafo regular, onde o diâmetro é proporcional ao tamanho da rede, ele pode ser transformado em uma rede de mundo pequeno, onde o número médio de arestas entre quaisquer par de vértices é muito pequeno, enquanto o coeficiente de agrupamento permanece grande. Portanto, o modelo de mundo pequeno está caracterizado por dois parâmetros importantes: um *coeficiente de agrupamento alto* e um *comprimento do caminho médio pequeno*.

Capítulo 3

Coleta, Tratamento e Visualização dos Dados

Este capítulo detalha cada um dos passos que foram executados nas fases de coleta, tratamento e visualização dos dados usados para análise das redes de colaboração abordadas nesta dissertação.

3.1 Dados do Estudo

Os dados usados neste estudo referem-se a pesquisadores associados a 48 instituições da América Latina (Tabela 3.1) com programas de Pós-graduação em Ciência da Computação e áreas afins relativos a um período de 20 anos, de 1994 até 2013, coletados da DBLP. O critério utilizado para selecionar as instituições considerou o total de publicações de cada pesquisador associado a essas instituições disponíveis na DBLP. Desta forma, foram consideradas aquelas instituições que tivessem um programa de Pós-graduação em Ciência da Computação com um corpo docente composto pelo menos por cinco pesquisadores com entradas na DBLP.

Segundo Franceschet [2011], a DBLP é internacionalmente respeitada pela qualidade dos seus dados. Hoje em dia, ela contém mais de 1,6 milhões de entradas referentes a artigos publicados em revistas científicas e anais de conferências das diversas subáreas da Ciência da Computação, bem como de áreas a fins. De outra parte, uma das principais vantagens da DBLP é que ela fornece os dados de cada pesquisador estruturados em XML (*eXtensible Markup Language*), permitindo assim a extração precisa dos atributos bibliográficos que descrevem cada publicação.

Pais	Instituição	# de Pub.	Taxa Pub. λ_i
Argentina	Universidad de Buenos Aires (UBA)	741	2,25
	Universidad Nacional de La Plata (UNLP)	373	2,80
	Universidad Nacional del Centro de la Provincia de Buenos Aires (UNICEN)	216	2,82
	Universidad Nacional del Sur (UNS)	214	2,05
Brasil	Universidade Federal do Rio de Janeiro (UFRJ-COPPE)	1.550	3,44
	Universidade Federal de Minas Gerais (UFMG)	1.601	4,17
	Universidade Federal do Rio Grande do Sul (UFRGS)	2.314	3,44
	Pontifícia Universidade Católica do Rio de Janeiro (PUC-RIO)	1.401	4,14
	Universidade Estadual de Campinas (UNICAMP)	1.516	3,04
	Universidade Federal de Pernambuco (UFPE)	2.054	3,31
	Universidade de São Paulo (USP)	933	2,61
	Universidade de São Paulo, São Carlos (USP-SC)	1.773	2,95
	Universidade Federal de Ceará (UFC)	315	2,06
	Universidade Federal do Amazonas (UFAM)	373	2,77
	Universidade Federal do Rio Grande do Norte (UFRN)	523	2,33
	Universidade Federal Fluminense (UFF)	881	2,30
	Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)	768	2,97
Chile	Pontifícia Universidad Católica de Chile (PUC-Chile)	367	2,52
	Universidad de Chile (UCHILE)	1.384	5,48
	Universidad Santiago de Chile (USACH)	219	1,70
	Universidad Técnica Federico Santa Maria (UTFSM)	495	1,68
	Universidad de Concepción (UDEC)	174	1,76
	Universidad de Valparaíso (UCV)	160	2,20
Colômbia	Universidad Icesi (ICESI)	35	0,75
	Pontificia Universidad Javeriana, Cali (PUJ-Cali)	38	1,36
	Universidad de los Andes (UNIANDES)	83	1,23
	Universidad del Valle (UNIVALLE)	32	0,87
	Universidad Nacional de Colombia, Medellín (UNAL-MED)	166	1,29
	Universidad Nacional de Colombia, Bogotá (UNAL-BOG)	112	1,76
Costa Rica	Universidad de Costa Rica (UCR)	73	0,89
Cuba	Universidad de La Habana (UH)	35	0,72
	Universidad de las Ciencias Informáticas (UCI)	26	0,69
	Universidad de Oriente (UO)	89	1,36
México	Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (CINVES-TAV)	618	3,88
	Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM)	155	1,66
	Universidad Autónoma del Estado de México (UAEMEX)	34	0,94
	Universidad Nacional Autónoma de México (UNAM)	919	2,73
	Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE)	426	3,14
	Instituto Tecnológico Autónomo de México (ITAM)	135	1,83
	Universidad de las Américas Puebla (UDLAP)	91	1,64
Paraguai	Universidad Nacional de Asunción (UNA)	84	1,49
Peru	Universidad Católica San Pablo (USCP)	41	0,90
	Pontificia Universidad Católica del Perú (PUCP)	46	1,05
Uruguai	Universidad de la República (UDELAR)	208	1,42
Venezuela	Universidad Central de Venezuela (UCV)	175	1,55
	Universidad de Carabobo (UC)	42	0,97
	Universidad Simón Bolivar (USB)	244	2,16
	Universidad de Los Andes - Mérida (ULA)	79	1,73

Tabela 3.1: Produção científica dos grupos de pesquisa das 48 instituições da América Latina no período 1994-2013.

A Tabela 3.1 apresenta a lista das instituições consideradas no estudo, com as respectivas taxas de publicação λ_i definidas pela Equação 3.1 abaixo,

$$\lambda_i = \frac{\sum_{y=1994}^{2013} \frac{P_{iy}}{R_{iy}}}{20} \tag{3.1}$$

onde P_{iy} é o total de publicações de cada instituição i no ano y e R_{iy} é o total de pesquisadores pertencentes à instituição i no ano y.

Uma vez definidos os grupos de pesquisa que foram selecionados para o estudo, a Figura 3.1 ilustra cada uma das fases do processo de coleta, tratamento e visualização dos dados.

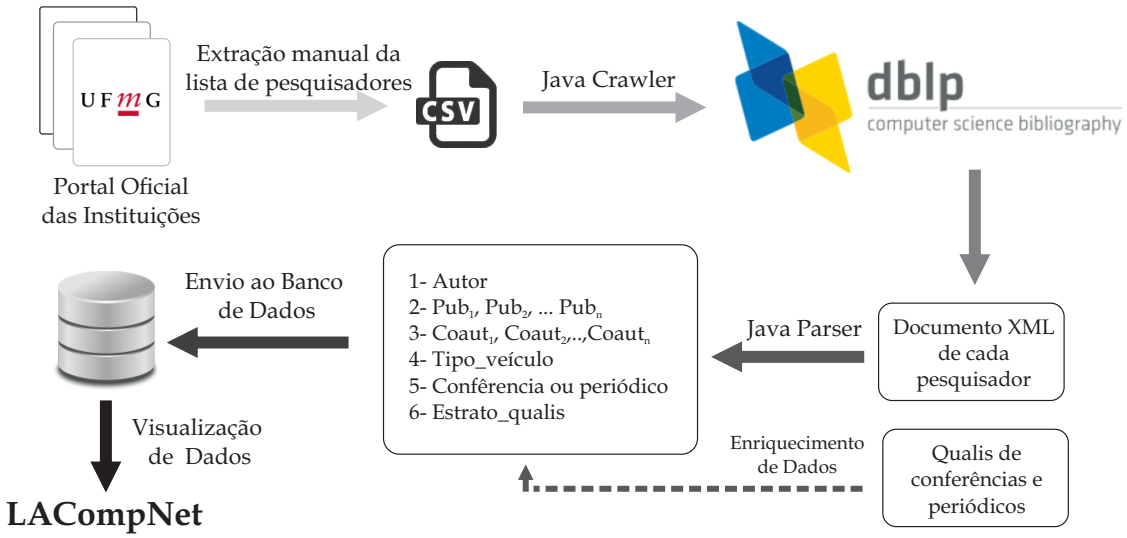


Figura 3.1: Fases do processo de coleta, tratamento e visualização de dados.

3.2 Coleta dos Dados

3.2.1 Coleta Manual

Nesta primeira fase, foram coletados manualmente os nomes dos pesquisadores pertencentes às instituições da América Latina[1]; a busca centrou-se em instituições com programas de pós-graduação em Ciência da Computação, Engenharia de Sistemas e Informática, denominações mais encontradas nos diversos países [Cuadros-Vargas et al., 2013]. Uma vez selecionados os pesquisadores, procedeu-se uma busca manual para

[1]Conforme informação existente na página oficial da instituição na data da coleta.

determinar aqueles que possuíam uma entrada na DBLP. Para cada um desses pesquisadores copiou-se o se nome como aparece na DBLP e o trecho da URL (do inglês *Uniform Resource Locator*) de sua página que o identifica unicamente. A Figura 3.2 ilustra um exemplo de um pesquisador e sua URL no formato da DBLP, neste caso o id do pesquisador é "m/Meira_Jr=:Wagner". Uma vez verificada a existência de

Figura 3.2: Página do pesquisador Wagner Meira Jr. na DBLP.

uma página para o pesquisador, foi gerado um arquivo de texto plano em formato CSV (do inglês *comma separated values*) com os seguintes dados: nome do pesquisador, instituição, URL, sigla da instituição, país e área de pesquisa.

3.2.2 Coleta das Páginas dos Pesquisadores

O arquivo CSV previamente gerado foi usado como entrada para o *crawler* para se fazer a extração automática da página em formato XML de cada pesquisador por meio de um dos servidores disponíveis na DBLP, da seguinte forma: A URL específica para extrair cada uma das páginas dos pesquisadores em formato XML encontra-se em `http://dblp.uni-trier.de/pers/xx/id_pesquisador` do servidor Trier 2; no exemplo anterior a URL utilizada para a extração do documento XML do pesquisador Wagner Meira Jr. foi `http://dblp.uni-trier.de/pers/xx/m/Meira_Jr=:Wagner`

como ilustrado na Figura 3.3. Usando-se a biblioteca Java *Simple API for XML* do projeto SAX[2], foram extraídos de cada uma das paginas DBLP em formato XML dos pesquisadores os seguintes atributos: número de artigos, nome do pesquisador na DBLP, tipo da publicação, lista de coautores, título da publicação, ano de publicação, título do veículo e URL do artigo.

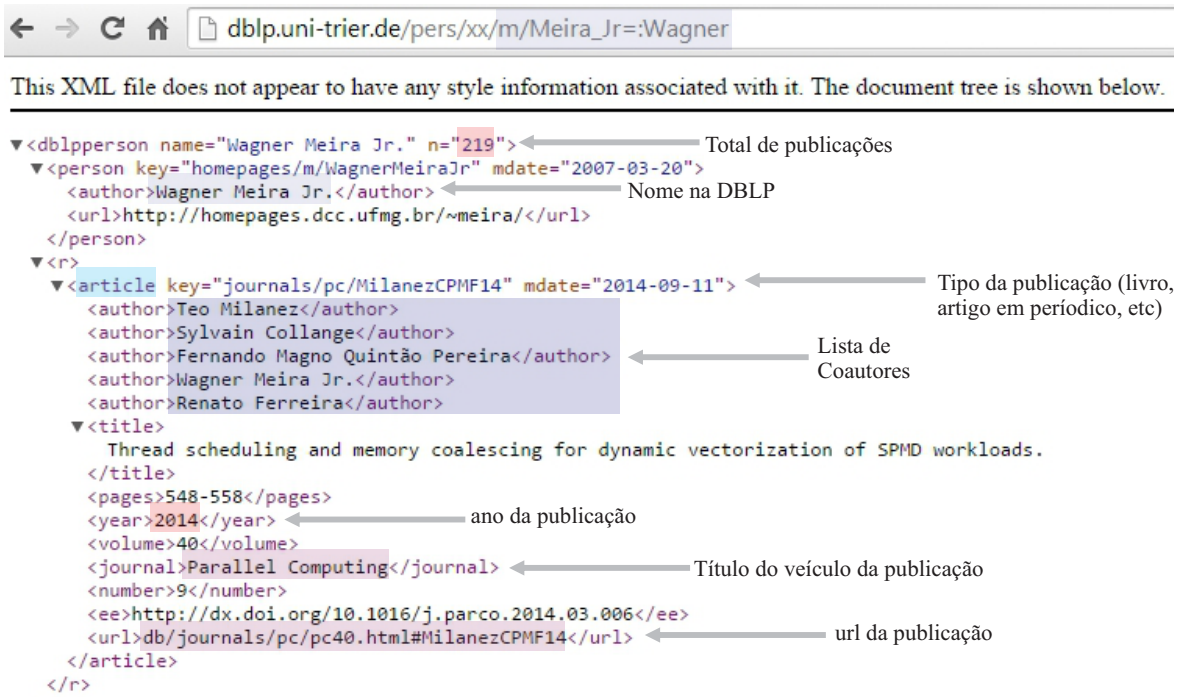


Figura 3.3: Trecho do documento XML gerado pela DBLP para o pesquisador Wagner Meira Jr.

3.3 Tratamento dos Dados

3.3.1 Extração dos Dados

Nesta fase, foi analisado cada um dos atributos extraídos das páginas coletadas e feita a limpeza dos dados através do uso de expressões regulares que permitiram retirar símbolos e caracteres especiais que poderiam posteriormente causar problemas ao se consultar o banco de dados. Por exemplo, havia nomes de pesquisadores contendo caracteres especiais como "Nicolás D'Ippolito" e que ao serem armazenados poderiam causar erros.

De outro lado, os dados foram transformados em objetos Java, onde cada um dos pesquisadores tivesse seus respetivos atributos, a fim de agilizar o tratamento dos

[2] `http://www.saxproject.org/`

mesmos. Adicionalmente, usando dados disponibilizados pela CAPES através do seu portal Web, foi feito o enriquecimento dos dados identificando a classificação Qualis dos veículos de cada uma das publicações, cobrindo-se um total de $65,27\%$ dos veículos encontrados na DBLP, dado que essa classificação só considera conferências e periódicos onde os pesquisadores brasileiros tenham previamente publicado. Mesmo assim, considerando o total de publicações, há uma cobertura de quase 85% em relação à classificação Qualis. Neste contexto, cada um dos veículos foi classificado de acordo com os estratos A1, A2, B1, B2, B3, B4 e B5, onde A1 é o estrato mais alto e B5 o mais baixo. No Capítulo 4 são usados esses estratos para se comparar a produção dos grupos de pesquisa gerados a partir das colaborações interinstitucionais.

3.3.2 Armazenamento dos Dados

Finalmente, nesta última fase de tratamento dos dados, eles foram armazenados em um banco de dados relacional usando o SGBD MySQL para facilitar a geração de consultas e permitir posteriores análises estatísticas. A modelagem do banco de dados foi realizada levando em consideração que as análises realizadas neste estudo têm por base os pesquisadores e suas publicações. Portanto, o esquema ilustrado na Figura 3.4 define 10 tabelas, onde duas delas se destacam pois armazenam os dados dos pesquisadores (tabela *"author"*) e das publicações (tabela *"publications"*) coletados da DBLP. A tabela *"authorpublication"* estabelece o relacionamento entre essas duas tabelas, o que permite identificar os autores de cada publicação.

Além disso, existe uma série de tabelas não menos importantes que enriquecem o banco de dados, a saber: a tabela *"venue"* armazena os dados dos veículos de publicação, e está diretamente relacionada com as tabelas *"venueType"*, que contém os tipos de veículo registrados na DBLP (*proceedings*, *inproceedings*, *articles* e *books*) e a tabela *"qualis"* que contém os sete estratos considerados pela CAPES para a avaliação das conferências e periódicos. A tabela *"institution"* armazena os nomes das instituições envolvidas no estudo e se encontra relacionada com as tabelas *"country"*, para conhecer o país ao qual pertence a instituição, e *"author"*, para determinar a afiliação de cada um dos pesquisadores. As últimas duas tabelas, *"researcharea"* e *"nameauthors"* contêm respetivamente as linhas de pesquisa dos autores e todos os seus possíveis registrados na DBLP.

Finalmente, vale ressaltar que todas as tabelas encontram-se devidamente normalizadas com o intuito de facilitar o processamento das consultas SQL, além de não

gerar inconsistências nos dados.

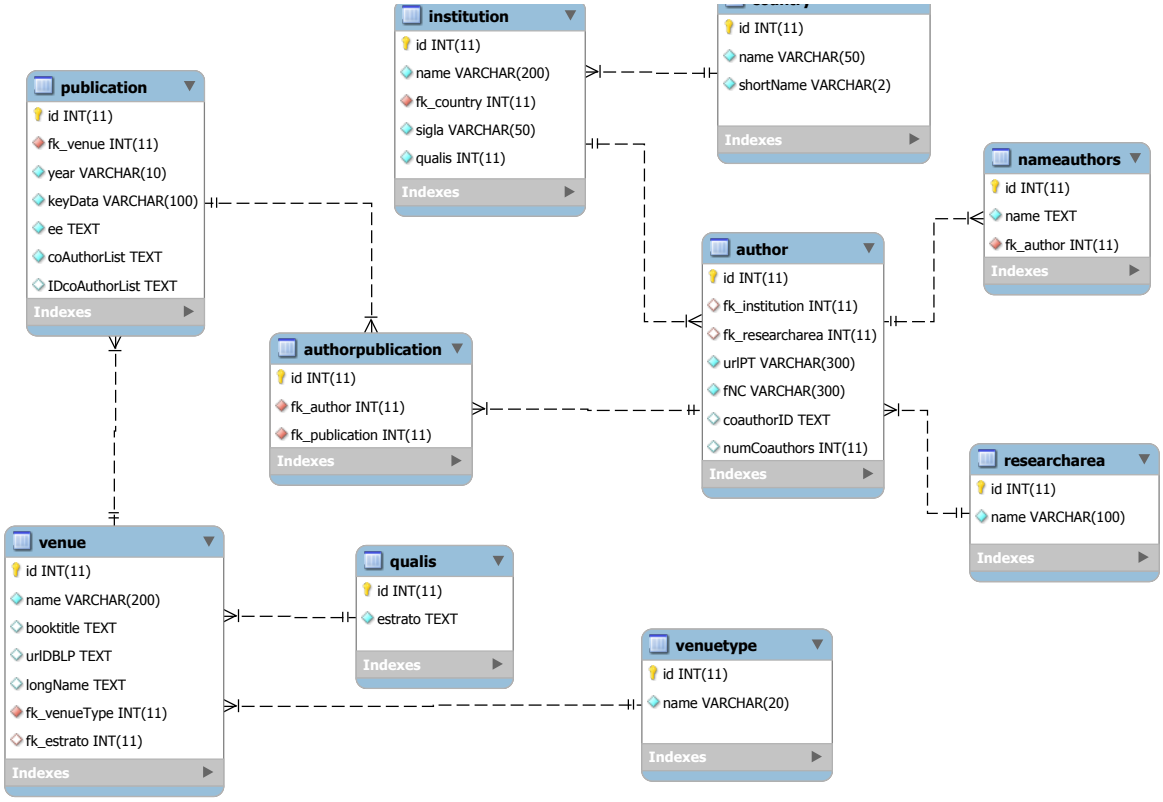


Figura 3.4: Esquema relacional do banco de dados.

Como exemplo, o Código 3.1 a seguir descreve a consulta SQL que seleciona os pesquisadores com mais de 50 publicações no período 1994-2003.

Código 3.1: **Consulta SQL**

```
1 SELECT author.fNC AS nome_Autor,
2        COUNT(DISTINCTROW(authorpublication.fk_publication)) AS Total,
3        institution.name AS Instituição
4 FROM authorpublication
5 INNER JOIN publication ON publication.id=authorpublication.fk_publication
6 INNER JOIN author ON author.id = authorpublication.fk_author
7 INNER JOIN institution ON institution.id = author.fk_institution
8 WHERE publication.year BETWEEN 1994 AND 2003
9 GROUP BY nome_Autor HAVING Total > 50
```

3.4 Visualização dos Dados: A Plataforma LACompNet

Nesta seção é apresentada a plataforma desenvolvida que suporta as análises feitas sobre os dados. A ferramenta foi desenhada a fim de ilustrar visualmente padrões frequentes, estatísticas e o perfil de publicação de cada uma das instituições, assim como as redes de colaboração entre instituições e países. Dentre as características mais importantes da ferramenta estão:

- Visualização da rede de colaboração composta pelas 48 instituições que formam parte do estudo, fornecendo métricas e propriedades da rede;
- Visualização das redes de colaboração entre países e instituições;
- Visualização da evolução temporal das publicações de cada um das instituições;
- Comparação da produção científica entre países e instituições.

A Figura 3.5 ilustra a tela inicial da plataforma que apresenta várias estatísticas e opções de navegação web.

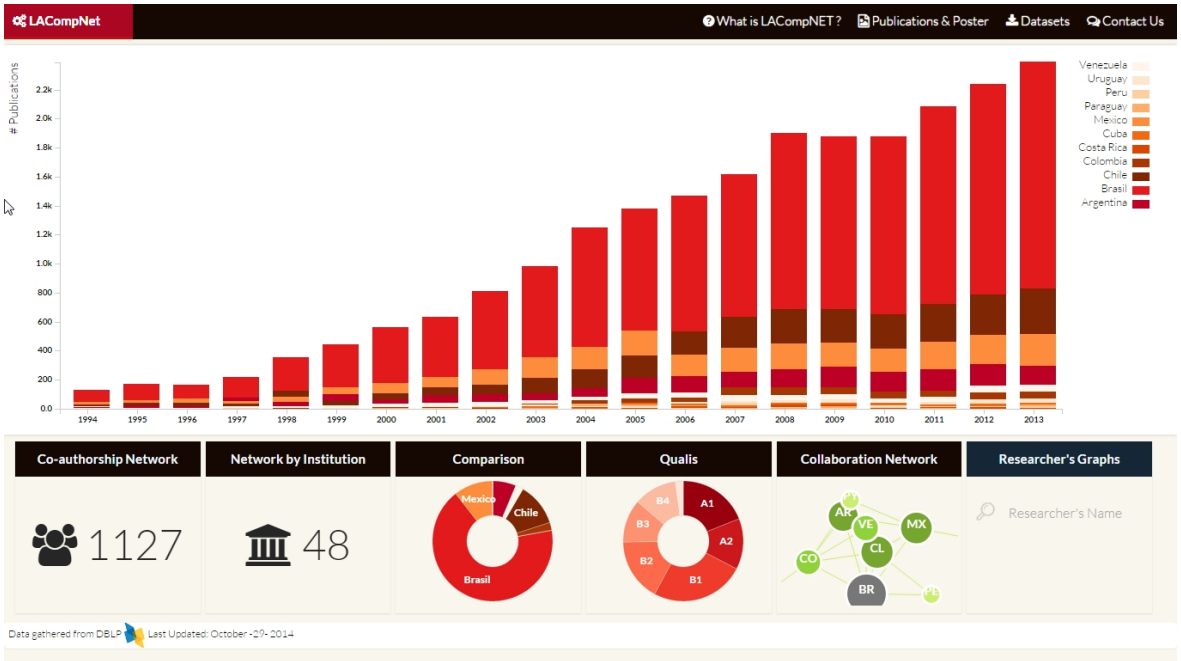


Figura 3.5: Tela inicial da plataforma LACompNet.

Para o desenvolvimento dessa ferramenta foram usadas diferentes linguagens e pacotes

Web, além do SGBD MySQL[3] e do servidor Web Apache[4] para o correto funcionamento da plataforma. O desenvolvimento Web foi feito na linguagem PHP como ferramenta de programação no lado do servidor, permitindo a geração do conteúdo dinâmico através da comunicação com o banco de dados. Para os cálculos das métricas de redes complexas usou-se Python por meio do pacote NetworkX[5], bem como JQuery[6] e JavaScript para habilitar o uso de bibliotecas gráficas como D3[7] e Sigma.js[8] para a visualização das redes e dos gráficos estatísticos.

[3] http://www.mysql.com/
[4] http://www.apache.org/
[5] https://networkx.github.io/
[6] http://jquery.com/
[7] http://d3js.org/
[8] http://sigmajs.org/

Capítulo 4

Redes de Colaboração

Este capítulo detalha como estão formadas as redes de colaboração em cada um dos países abordados neste estudo. Especificamente, a Seção 4.1 apresenta o perfil de publicação de cada um dos países, e a Seção 4.2 apresenta uma análise das redes de colaboração científica e como os pesquisadores têm interagido ao longo do tempo.

4.1 Estatísticas Gerais

A fim de caracterizar a produção científica dos países latino-americanos envolvidos neste estudo, a Tabela 4.1 apresenta as estatísticas gerais ao longo do período 1994-2013.

Estatística	**LACompNet**
Total de pesquisadores	**18.523**
De instituições latino-americanas	1.127
Outros	17.396
Total de publicações	**22.329**
Média de publicações por pesquisador	27,12 (36,25†)
Média de autores por publicação	3,49 (1,86†)
Média de colaboradores por pesquisador	28,77 (33,18†)

† Valor do desvio padrão

Tabela 4.1: Estatísticas gerais sobre o total das publicações dos pesquisadores Latino-Americanos no período 1994-2013.

De acordo com esses dados, nota-se que o número de pesquisadores vinculados às instituições latino-americanas corresponde a $6,08\%$ do total de autores. A média geral

de publicações por pesquisador mostra um claro indício que há um bom número de pesquisadores com uma alta taxa de publicações durante o período 1994-2013, o que é confirmado pelos pesquisadores com maior número de publicações durante o período, conforme a Tabela 4.2.

Pesquisador	**Instituição**	**# Publicações**
Gonzalo Navarro	UCHILE	336
Ricardo A. Baeza-Yates	UCHILE	299
Carlos A. Coello Coello	CINVESTAV	276
Carlos José Pereira de Lucena	PUC-RIO	252
Luigi Carro	UFRGS	247
Jano Moreira de Souza	UFRJ-COPPE	200
Wagner Meira Jr.	UFMG	199
André Carlos Ponce Leon Ferreira de Carvalho	USP-ICMC	196
Marcos André Gonçalves	UFMG	195
Antonio A. F. Loureiro	UFMG	194

Tabela 4.2: Classificação dos 10 pesquisadores com maior volume de publicações no período 1994-2013.

Por outro lado, 4,88% dos pesquisadores tiveram mais de 100 publicações durante o período, gerando valores extremos que afetam diretamente a média geral de publicações, o que é confirmado pelo alto desvio padrão de 36,25. A média geral de autores por publicação observada, se comparada com outras redes de colaboração científica estudadas [Albert & Barabási, 2002; Newman, 2004, 2001a], é muito próxima aos valores das áreas da Física, Biologia, Biomedicina e Astrofísica que estão entre 3,35 e 3,75 colaboradores por publicação.

Com relação à média de colaboradores por pesquisador pode-se observar que há uma variação também maior do que a média geral, dado o alto valor do desvio padrão de 33,18. A Tabela 4.3 mostra os dez pesquisadores com maior número de colaboradores. Do mesmo modo, podemos conhecer os autores que têm maior cooperação com pesquisadores de outras instituições da América Latina como mostrado na Tabela 4.4.

Pesquisador	Instituição	# Colaboradores
Wagner Meira Jr.	UFMG	241
Ricardo A. Baeza-Yates	UCHILE	235
Carlos José Pereira de Lucena	PUC-RIO	205
Marcos André Gonçalves	UFMG	202
Luigi Carro	UFRGS	191
Jano Moreira de Souza	UFRJ-COPPE	175
Alessandro Garcia	PUC-RIO	172
Silvio Romero de Lemos Meira	UFPE	169
Carlos A. Coello Coello	CINVESTAV	166
Jussara M. Almeida	UFMG	164

Tabela 4.3: Classificação dos 10 pesquisadores com maior número de colaboradores no período 1994-2013.

Pesquisador	Instituição	# Colaboradores
Thaís Vasconcelos Batista	UFRN	23
Carlos José Pereira de Lucena	PUC-RIO	22
Ricardo A. Baeza-Yates	UCHILE	18
Julio Cesar Sampaio do Prado Leite	PUC-RIO	17
Altigran Soares da Silva	UFAM	16
Alberto H. F. Laender	UFMG	15
Uirá Kulesza	UFRN	15
Fernando Castor	UFPE	15
Jayme Luiz Szwarcfiter	UFRJ-COPPE	14
Fábio Protti	UFF	14

Tabela 4.4: Classificação dos 10 pesquisadores com maior número de colaboradores de outras instituições da América Latina no período 1994-2013.

4.1.1 Média Geral de Publicações

Considerando os dados do Capítulo 3, a Tabela 4.5 apresenta um resumo da produção científica em Ciência da Computação da América Latina no período 1994-2013, a qual pode ser dividida em quatro grupos de acordo a média geral $\bar{X}_c$.

Tomando como base a Equação 3.1 do Capítulo 3, a média geral de publicações por país $\bar{X}_c$ foi calculada de acordo a Equação 4.1

$$\bar{X}_c = \frac{\sum\limits_{i \in U_c} \lambda i}{\mid U_c \mid} \tag{4.1}$$

onde U_c é o conjunto total de instituições i do pais c.

País	$\bar{X}_c$
Brasil	3,04
Chile	2,56
Argentina	2,48
México	2,26
Venezuela	1,60
Paraguai	1,49
Uruguai	1,42
Colômbia	1,21
Peru	0,98
Cuba	0,92
Costa Rica	0,89

Tabela 4.5: Produção média por pesquisador por país no período 1994-2013.

De acordo aos dados da Tabela 4.5, a produção cientifica dos países latino-americanos pode ser divida assim: o primeiro grupo é formado pelo Brasil com uma média geral superior a mais de três publicações por pesquisador nas duas décadas, o segundo por Argentina, Chile e México com uma média geral entre duas e três publicações nas duas décadas, o terceiro por Colômbia, Paraguai, Uruguai e Venezuela com uma média geral de mais de uma publicação e menos de duas publicações nas duas décadas e o último por Costa Rica, Cuba e Peru com menos de uma publicação nas duas décadas.

Neste ponto podemos analisar como foi o incremento na média geral de publicações de uma década para outra. A Tabela 4.6 mostra o total de publicações por década e o incremento em cada um dos países.

O gráfico da Figura 4.1 ilustra a evolução temporal da média de produção por país ano a ano. Note-se que o maior incremento na média geral de produção $\bar{X}_c$ está em países como Costa Rica, Cuba e Peru com pouca tradição na área de Ciência da Computação; no entanto, na última década esses países conseguiram um incremento notório dado o estabelecimento de novas colaborações inclusive com pesquisadores de outros países, como descrito na Subseção 4.1.2. A seguir aparecem Colômbia e Paraguai que também apresentaram um crescimento expressivo de sua produção na ultima década. Em contraste, países com uma maior experiência e grupos já estabelecidos

País	**Total Pesquisadores**		**Média Geral $\bar{X}_c$**		**Inc. $\bar{X}_c$**
	1994-2003	2004-2013	1994-2003	2004-2013	
Argentina	57	100	1,59	3,37	111,95 %
Brasil	397	495	1,97	4,12	109,14 %
Chile	74	145	1,52	3,60	136,84 %
Colômbia	13 (5)	60	0,60	1,90	216,67 %
Costa Rica	2	15	0,20	1,58	**690,00 %**
Cuba	11	31	0,27	1,39	**414,82 %**
México	76	117	1,44	3,08	113,89 %
Paraguai	2	8	0,70	2,27	224,28 %
Peru	4	17	0,35	1,60	**357,14 %**
Uruguai	12	41	0,96	1,87	94,79 %
Venezuela	41	64	1,21	2,00	65,29 %

Tabela 4.6: Incremento na média geral de publicações nas duas décadas por país.

apresentaram um incremento estável como nos casos do Brasil, Argentina, Chile e México. Por outro lado, Uruguai e Venezuela têm apresentado um crescimento mais lento ao longo dos últimos anos, não obstante com uma média geral de publicações superior ou similar a países como Colômbia e Paraguai.

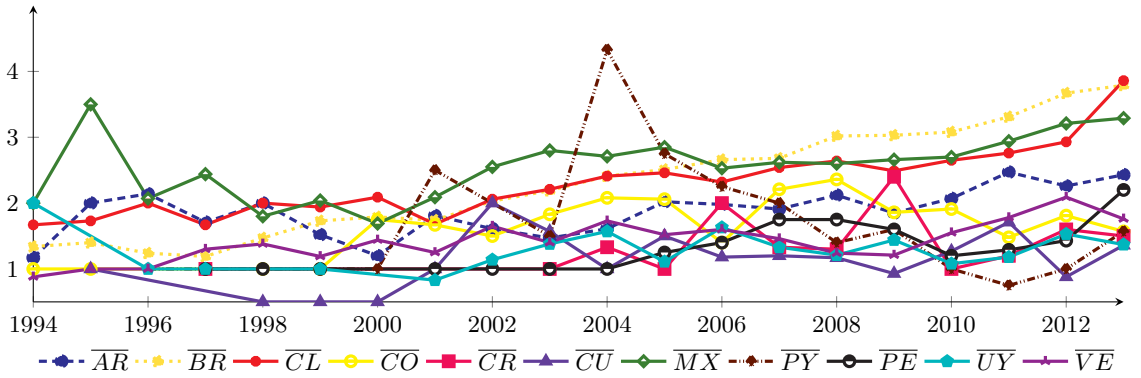


Figura 4.1: Média geral de publicações ano a ano por país no período 1994-2013.

O gráfico da Figura 4.2 ilustra a distribuição da produção científica de cada um dos países de acordo com a sua taxa de publicação. A seguir, apresentamos uma breve análise dessa produção.

A **Argentina**, representada por quatro instituições (UBA, UNLP, UNICEN e UNS), apresenta uma taxa de publicação por instituição entre 2,05 e 2,82, valores correspondentes aos grupos da UNS e UNICEN respectivamente; os valores mínimo e

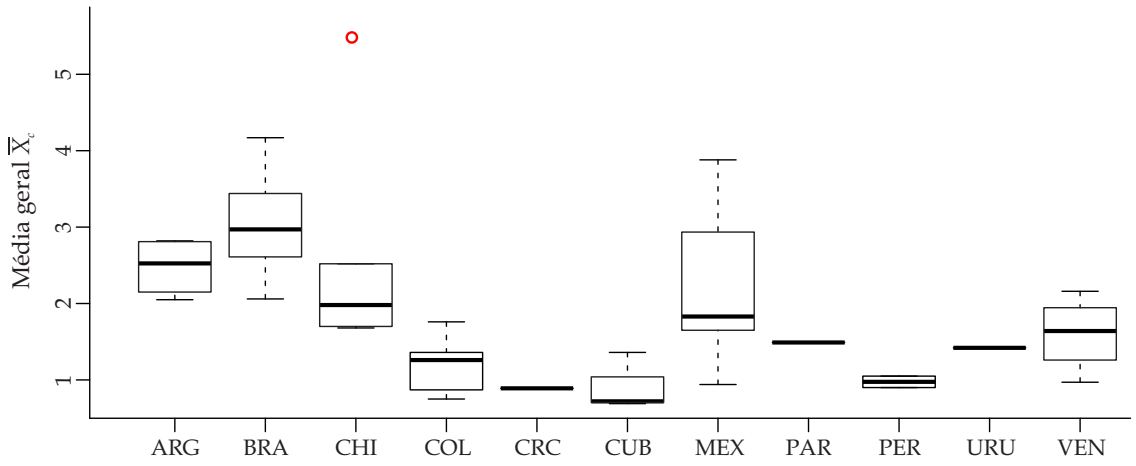


Figura 4.2: Distribuição de publicações por país no período 1994-2013.

máximo estão muito perto do primeiro quartil (2,20) e do terceiro quartil (2,81). Neste caso, a UBA possui uma taxa de publicação de 2,25 se posicionando muito perto da mediana 2.53, em contraste com a UNLP cuja taxa de 2,80 publicações coincide com a mediana. Neste contexto, pode-se dizer que as instituições da Argentina apresentam uma taxa de publicação muito estável e perto da média geral de 2,48 publicações com um desvio padrão de 0,39.

O **Brasil**, representado pelas treze instituições mas bem classificadas segundo a CAPES (UFRJ-COPPE, UFMG, UFRGS, PUC-RIO, UNICAMP, UFPE, USP, USP-SC, UFC, UFAM, UFRN, UFF e PUCRS), apresenta uma taxa de publicação por instituição entre 2,06 e 4,17, valores correspondentes à UFC e UFMG respectivamente. O valor da mediana é 2,97 correspondente à PUCRS. Neste contexto, as instituições USP-SC com uma taxa de publicação de 2,95, UFAM com uma taxa de publicação de 2,77 e USP com uma taxa de publicação de 2,61 se posicionam abaixo do valor da mediana, em contraste com as instituições UFRJ-COPPE, com uma taxa de publicação de 3,44, UFRGS, com uma taxa de publicação de 3,44, UNICAMP, com uma taxa de publicação de 3,04, e UFPE, com uma taxa de publicação de 3,31 que se posicionam acima da mediana.

Por outro lado, três instituições apresentam valores atípicos: UFC, UFF e UFRN, com taxas de publicação de 2,06, 2,30 e 2,33 respectivamente. Adicionalmente

duas instituições apresentam valores extremos muito altos em relação à mediana com valores de 4,14 e 4,17 para PUC-RIO e UFMG, respectivamente. Considerando os valores apresentados, podemos concluir que o Brasil apresenta uma dispersão dos seus valores em termos de taxa de publicação das instituições ao longo dos 20 anos, que é afetada por valores extremos, o que pode ser evidenciado por uma média geral de 3,04 e um desvio padrão de 0,66.

O **Chile**, representado por seis instituições (PUC-Chile, UCHILE, USACH, UTFSM, UDEC e UCV), apresenta uma taxa de publicação por instituição que está entre 1,68 e 5,48, correspondentes à UTFSM e UCHILE respectivamente. No caso do Chile, encontramos um valor atípico da taxa de publicação correspondente à UCHILE, que visivelmente está fora da distribuição dos dados, já que dois de seus pesquisadores, Gonzalo Navarro e Ricardo Baeza-Yates, apresentam altos índices de produção científica nas duas décadas. Por outro lado, a USACH com uma taxa de publicação de 1,70 se posiciona abaixo da mediana, em contraste com a UCV com uma taxa de publicação de 2,20 acima da mediana. Duas outras instituições do Chile apresentam valores extremos para a taxa de publicação, UDEC com 1,76 e PUC-Chile com 2,52 publicações por pesquisador. Neste caso, pode-se observar que a taxa média geral de publicação no Chile foi de 2,56 com um desvio padrão de 1,47 claramente afetado pela produção da UCHILE.

A **Colômbia**, representada por seis instituições (ICESI, PUJ-Cali, UNIANDES, UNIVALLE, UNAL-MED e UNAL-BOG), apresenta uma taxa de publicação por instituição que está entre 0,75 e 1,76, correspondentes à ICESI e UNAL-BOG, respectivamente. Neste caso, UNIANDES encontra-se no primeiro quartil com uma taxa de publicação de 1,23 e a UNAL-MED com 1,29 se posicionando no terceiro quartil, ambos os casos próximos ao valor da mediana de 1,26. Os valores extremos são apresentados pela UNIVALLE com 0,87 e PUJ-Cali com 1,36. A taxa média geral de publicações foi de 1,21 com um desvio padrão de 0,36. Note-se que há dois grupos (UNAL-BOG e UNAL-MED), que sendo de diferentes cidades pertencem à mesma instituição nacional, a UNAL, com quase $59,67\%$ da produção total das publicações na Colômbia durante o período 1994-2013.

Cuba, representada por três instituições (UH, UCI e UO), apresenta uma taxa de publicação que está entre 0,69 e 1,36, com uma mediana de 0,72 correspondente à UH, sendo que a UCI encontra-se abaixo desse valor e a UO apresenta o valor extremo de 1,36 publicações por pesquisador no período. Em Cuba, a

taxa média geral foi de 0,92 com um desvio padrão de 0,38 mostrando uma dispersão dos valores dada pela Universidad de Oriente (UO) que situa-se longe da média.

O **México**, representado por sete instituições (CINVES-TAV, ITESM, UAEMEX, UNAM, CICESE, ITAM e UDLAP), apresenta uma taxa de publicação que está entre 0,94 e 2,89, correspondentes à UAMEX e CINVES-TAV respectivamente. A mediana situa-se em torno de 1,83, valor no qual está posicionada a ITAM. Neste contexto, no primeiro quartil (25%) encontra-se o ITESM com uma taxa de 1,66, enquanto que a UNAM com uma taxa de 2,73 encontra-se no terceiro quartil. As outras instituições situam-se nos extremos da distribuição dos dados, a UDLAP com 1,64, em contraste com CICESE com 3,14. A média geral foi de 2,26 com um desvio padrão de 1,02 evidenciando uma clara dispersão dos valores.

A **Venezuela**, representada por quatro instituições (UCV, UC, USB e ULA), apresenta uma taxa de publicação entre 0,97 e 2,16, correspondentes à UC e USB respectivamente. O valor da mediana é de 1,64 publicações por pesquisador. A UCV possui uma taxa de 1,55, portanto abaixo da mediana, em contraste com a ULA que possui uma taxa de 1,73. A média geral foi de 1,60 com um desvio padrão de 0,49 representando uma dispersão dos dados causada pelos valores extremos da UC e USB respectivamente.

Finalmente, Costa Rica e Peru, representados por uma (UCR) e duas instituições (USCP e PUCP) respectivamente, possuem os grupos com menor produção científica ao longo dos 20 anos, com uma taxa de publicação inferior a uma publicação por pesquisador, em contraste com Paraguai e Uruguai, representados por uma única instituição que, em ambos os casos, apresentaram uma taxa de publicação superior à da Colômbia e muito similar à da Venezuela, países com mais de três instituições.

4.1.2 Formação de Colaborações Internacionais

Com o intuito de conhecer as colaborações entre países a partir das publicações científicas, foi realizado um processo de mineração de dados sobre o total de publicações conjuntas, em periódicos ou conferências, que envolvessem duas ou mais instituições em pelo menos dois anos do período estudado (1994-2013). Para isto, foi necessário extrair do banco de dados um subconjunto de dados correspondente às publicações referentes ao período 1994-2013 por meio de uma consulta SQL (Código 4.1) envolvendo os seguintes atributos:

- Identificador da publicação, Qualis, Ano da publicação, Veículo, Tipo do Veículo, Instituição$_1$, $\cdots$, Instituição$_n$, País$_1$, $\cdots$, País$_n$, Autor$_1$, $\cdots$, Autor$_n$.

Código 4.1: **Extração dos dados**

```
1  SET SESSION group_concat_max_len = 10240000;
2  SELECT publication.id AS idPub, qualis.estrato AS qualis, publication.
       keyData AS publication, publication.year AS pubYear, venue.urlDBLP AS
       venue, venuetype.name AS venueType, GROUP_CONCAT(author.fNC SEPARATOR
       ",") AS author, GROUP_CONCAT (DISTINCTROW(institution.name) SEPARATOR
       ",") AS institution, GROUP_CONCAT (DISTINCTROW(country.name) SEPARATOR
        ",") AS country
3  FROM authorpublication
4  INNER JOIN publication ON publication.id = authorpublication.
       fk_publication
5  INNER JOIN author ON author.id = authorpublication.fk_author
6  INNER JOIN institution ON institution.id = author.fk_institution
7  INNER JOIN country ON country.id = institution.fk_country
8  INNER JOIN venue ON venue.id = publication.fk_venue
9  INNER JOIN venuetype ON venuetype.id = venue.fk_venuetype
10 INNER JOIN qualis ON qualis.id = venue.fk_estrato
11 WHERE publication.year BETWEEN 1994 AND 2013
12 GROUP BY idPub
```

Os dados resultantes referem-se a 22.329 publicações, 18.523 autores, 2.312 veículos, 5 tipos de veículo e 48 instituições nas duas décadas. Com base nesses dados, foi aplicado o algoritmo *Apriori* [Zaki & Meira Jr, 2014] a fim de determinar padrões frequentes de publicação entre países. Adicionalmente, com o intuito de analisar a qualidade das publicações de cada um dos países envolvidos no estudo, o nosso banco de dados foi enriquecido com a informação do Qualis das conferências[1] e periódicos[2] avaliados pela CAPES, que foram obtidos através do sistema WebQualis[3], como mencionado no Capítulo 3. As Tabelas 4.7 e 4.8 mostram o total de cooperações binacionais mineradas a partir dos períodos estabelecidos, junto com o estrato Qualis dos veículos onde as produções foram publicadas. Ressalte-se que não foi possível identificar durante esse período cooperações que envolvessem três ou mais países da América Latina.

Note-se que houve um incremento do 132% no total de cooperações de uma década para outra, de 75 colaborações no período 1994-2003 passou-se a 174 no

[1]Última atualização Agosto/31/2012. `http://www.capes.gov.br/images/stories/download/avaliacao/Comunicado_004_2012_Ciencia_da_Computacao.pdf`

[2]Última atualização Janeiro/9/2014

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

Colaboração	Classificação Qualis								Total
	A1	A2	B1	B2	B3	B4	B5	$SC^{\dagger}$	
Argentina - Brasil	8	3	1	11	3	2	3	5	36
Brasil - Chile	6	5	5	0	1	0	0	2	19
Brasil - Peru	0	0	1	0	1	1	0	-	3
Brasil - México	1	0	0	0	0	0	0	-	1
Brasil - Venezuela	1	1	0	0	0	0	0	-	2
Chile - Colômbia	0	0	1	0	0	0	0	-	1
Chile - México	3	1	6	1	1	0	1	-	13

† SC: Veículos sem classificação Qualis

Tabela 4.7: Distribuição de publicações em conjunto no período 1994-2003.

período 2004-2013. De fato, na primeira década (Tabela 4.7) foram identificadas sete cooperações concentradas em apenas quatro países, Argentina, Brasil, Chile e México. Por outro lado, na segunda década (Tabela 4.8) passou-se a 19 cooperações envolvendo 11 países cujas instituições possuem ampla história na América Latina.

Analisando as colaborações entre os grupos na primeira década podemos ressaltar que o maior número de publicações em conjunto envolveu pesquisadores da UFMG e UCHILE com $42,10\%$ do total de publicações e da UFMG, UFAM e UCHILE com $36,84\%$, entre as de maior destaque neste período. Quanto à colaboração entre Argentina e Brasil, o maior número de publicações foi entre os grupos UNLP e PUC-Rio com $97,22\%$ do total das publicações. Do mesmo modo, na colaboração entre Chile e México, o maior número de publicações ocorreu entre os grupos das CICESE e UCHILE com 100% das publicações. Ressalte-se, também, que o maior número de colaborações entre esses grupos ocorreu no estrato A1, demostrando a qualidade dessas colaborações.

Na segunda década, é evidente um fortalecimento nas colaborações já existentes entre os países com ampla história na área da Ciência da Computação na América Latina, além de uma diversificação de novas colaborações com instituições da Colômbia, Peru, México, Uruguai e Venezuela. Nessa década, note-se que o maior número de colaborações ocorreu no estrato B1, corroborando a qualidade das colaborações segundo a classificação Qualis. No caso da colaboração entre Argentina e Brasil houve um incremento de $36,11\%$. Já a colaboração entre Brasil e Chile teve um incremento de $52,63\%$ e a colaboração entre Chile e México um incremento de $138,46\%$. Neste contexto, analisando as colaborações entre Argentina e Brasil, o maior número de

Colaboração	Classificação Qualis								Total
	A1	A2	B1	B2	B3	B4	B5	$SC^{\dagger}$	
Argentina - Brasil	4	6	5	8	8	5	2	11	49
Argentina - Chile	1	1	3	3	1	1	0	2	12
Argentina - Colômbia	1	0	0	0	0	0	0	0	1
Argentina - México	0	0	0	1	0	0	0	0	1
Brasil - Chile	7	4	4	6	1	3	0	4	29
Brasil - Colômbia	1	0	0	1	1	0	2	2	7
Brasil - México	0	0	1	0	0	0	0	0	1
Brasil - Peru	1	0	6	0	0	0	1	1	9
Brasil - Uruguai	0	0	1	1	1	1	0	0	4
Brasil - Venezuela	1	0	0	0	0	1	0	1	3
Chile - Colômbia	0	0	1	0	2	1	0	0	4
Chile - México	5	5	3	4	2	1	2	9	31
Chile - Peru	0	0	1	0	1	0	0	2	4
Colômbia - Costa Rica	0	0	0	0	0	0	0	1	1
Colômbia - Paraguai	0	0	1	1	0	2	0	1	5
Colômbia - Venezuela	0	0	0	0	0	0	0	0	1
Costa Rica - Uruguai	0	0	0	0	0	1	0	0	1
Cuba - México	0	1	2	0	0	0	0	2	5
México - Paraguai	1	3	1	0	0	0	0	1	6

† SC: Veículos sem classificação Qualis

Tabela 4.8: Distribuição de publicações em conjunto no período 2004-2013.

cooperações ocorreu entre os grupos da UBA e UFRJ-COPPE, com $51,02\%$, e da UNLP e PUC-Rio, com $30,61\%$. Do mesmo modo, o maior número de colaborações entre a Argentina e Chile envolveu os grupos da UBA e USACH, com $33,33\%$, e da UNLP e UCHILE, com o $41,67\%$. Quanto à cooperação entre Brasil e Chile é destacável a parceria entre os grupos da UFMG e UCHILE que abrange o $68,97\%$ do total das publicações, motivada pelo número de pesquisadores das duas instituições que trabalham em conjunto. Igualmente, a colaboração entre o Brasil e Peru tem como destaque a parceria entre os grupos da USP-SC e USCP com $77,78\%$ do total das publicações. Já entre Chile e México, a maior colaboração ocorreu entre os grupos do CICESE e da UCHILE com $80,65\%$ das publicações. Em suma, é importante ressaltar o fato que as colaborações representam uma ferramenta chave na produção científica de qualidade. A Tabela 4.9 apresenta a classificação das 10 maiores colaborações entre grupos ao longo do tempo, realçando a ampla colaboração entre as instituições brasileiras.

Colaboração	Total de Colaborações
UFMG e UFAM	158
UFRJ-COPPE e UFF	158
PUC RS e UFRGS	97
PUC-CHILE e UCHILE	96
PUC-RIO e UFF	87
PUC-RIO e UFRN	78
UFMG e UNICAMP	75
IME-USP e UNICAMP	69
PUC-RIO e UNLP	50
ICMC-USP e UFPE	50

Tabela 4.9: Classificação das 10 colaborações entre instituições em termos de produtividade no período 1994-2013.

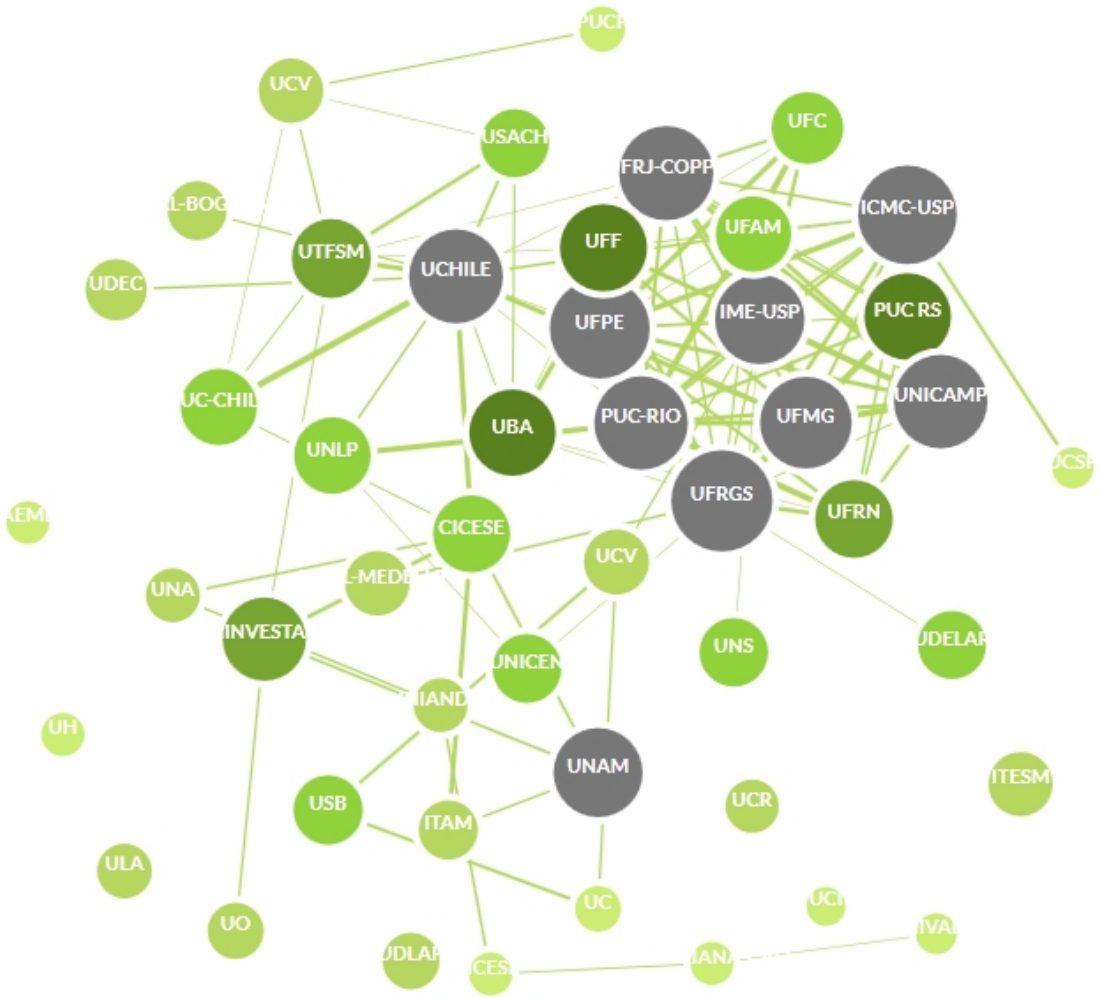


Figura 4.3: Grafo de colaboração entre as instituições no período 1994-2013.

A Figura 4.3 ilustra o grafo de colaboração entre os grupos de pesquisa latino-americanos no período 1994-2013. Nota-se que o grafo apresenta um agrupamento formado pelas instituições brasileiras, evidenciando um alto grau de proximidade e colaboração entre elas, confirmado também pelo alto coeficiente de agrupamento entre seus grupos que é na média de 0,59.

De fato, entre as dez instituições com maior valor da métrica de proximidade (ver Apêndice A), encontram-se nove brasileiras e a UCHILE, que é a instituição que possui o maior número de pesquisadores que colaboram com instituições brasileiras. Por outro lado, é de ressaltar que há cinco instituições que não possuem qualquer colaboração com as demais instituições consideradas neste estudo, a saber: ITESM, UAEMEX e UDLAP do México, UCI de Cuba e ULA da Venezuela. O Apêndice A apresenta as estatísticas gerais das redes de colaboração entre instituições latino-americanas no período 1994-2013.

4.1.3 Qualidade das Colaborações Segundo as Publicações Produzidas

Neste ponto, nosso interesse é efetuar uma análise qualitativa da produção científica dos grupos de pesquisa mineradas conforme descrito na Subsecção 4.1.2. Nesse contexto foram considerados dois índices, I_T[4] e I_S[5], para aferir a qualidade desses grupos, considerando tanto artigos em conferências como em periódicos. Sendo assim, os índices I_T e I_S foram definidos respectivamente pelas Equações 4.2 e 4.3 abaixo,

$$I_{T_i} = \frac{TP_{A1} + 0,85 \times TP_{A2} + 0,70 \times TP_{B1} + 0,55 \times TP_{B2} + 0,40 \times TP_{B3} + 0,25 \times TP_{B4} + 0,10 \times TP_{B5}}{\log\left(t_{a_i} \times t_{c_i}\right)} \tag{4.2}$$

$$I_{S_i} = \frac{TP_{A1} + 0,85 \times TP_{A2} + 0,70 \times TP_{B1}}{\log\left(t_{a_i} \times t_{c_i}\right)} \tag{4.3}$$

onde TP é o total de publicações de cada estrato, t_a é o total de autores[6] da colaboração i e t_c é o tempo da colaboração i definido em número de anos ocorridos desde o seu início.

As Tabelas 4.10 e 4.11 mostram as colaborações mineradas no período 1994-2013,

[4]Considera publicações em todos os estratos A1, A2, B1, B2, B3, B4 e B5.
[5]Considera apenas publicações nos estratos superiores A1, A2 e B1.
[6]Autores que pertencem às instituições latino-americanas consideradas neste estudo.

incluindo o total de autores envolvidos (t_a), o tempo total de anos de colaboração (t_c) e o total de publicações por estrato ($TP_{estrato}$), ordenadas respectivamente pelos índices I_T e I_S.

Colaboração	t_a	t_c	TP_{A1}	TP_{A2}	TP_{B1}	TP_{B2}	TP_{B3}	TP_{B4}	TP_{B5}	I_T
Argentina - Brasil	36	19	12	9	6	19	11	7	5	14,444
Brasil - Chile	29	19	13	9	9	6	2	3	0	11,601
Chile - México	14	15	8	6	9	5	3	1	3	10,292
Argentina - Chile	8	7	1	1	3	3	1	1	0	3,575
México - Paraguai	3	8	1	3	1	0	0	0	0	3,079
Colômbia - Paraguai	2	2	0	0	1	1	0	2	0	2,907
Brasil - Peru	13	15	1	0	7	0	1	1	1	2,904
Chile - Colômbia	4	10	0	0	2	0	2	1	9	2,091
Cuba - México	2	7	0	1	2	0	0	0	0	1,963
Brasil - Venezuela	5	8	2	1	0	0	0	1	0	1,935
Brasil - Uruguai	8	6	0	0	1	2	2	2	0	1,844
Brasil - Colômbia	7	9	1	0	0	1	1	0	2	1,195
Brasil - México	4	14	1	0	1	0	0	0	0	0,972
Chile - Peru	4	5	0	0	1	0	1	0	0	0,845

Tabela 4.10: Classificação das colaborações conforme o índice I_T no período 1994-2013.

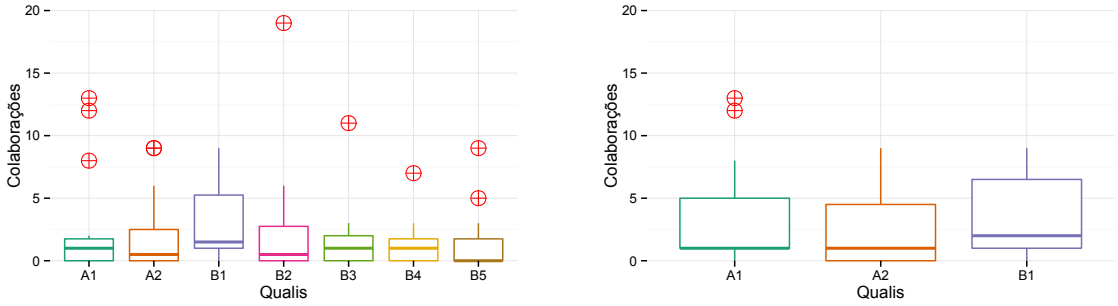
Na Tabela 4.10, ressalta-se que o maior número de cooperações envolve pesquisadores do Brasil, tendo como seus colaboradores pesquisadores da Argentina, Chile, Colômbia, México, Peru, Uruguai e Venezuela. Além disso, podemos ressaltar que países com menor tradição na área de Ciência da Computação, como Colômbia, Peru e Paraguai, tendem a publicar com países de maior tradição na região. Por outro lado, há uma maior diversificação das publicações em todos os estratos, com uma maior taxa de colaborações no estrato B1.

Na Tabela 4.11 podemos observar que nas colaborações entre países com maior taxa média de publicações há uma tendência que elas ocorram nos estratos superiores, como acontece com as colaborações entre Argentina - Brasil, Brasil - Chile e Chile - México.

As Figuras 4.4a e 4.4b ilustram a distribuição dos índices I_T e I_S respectivamente. Nota-se a presença de valores atípicos que em todos os casos, correspondem a colaborações entre os maiores produtores de artigos científicos da região, ou seja, Argentina, Brasil, Chile e México.

Colaboração	t_a	t_c	TP_{A1}	TP_{A2}	TP_{B1}	I_S
Brasil - Chile	25	19	13	9	9	10,068
Argentina - Brasil	21	19	12	9	6	9,170
Chile - México	12	15	8	6	9	8,602
Brasil - Peru	9	9	1	0	7	3,091
México - Paraguai	3	8	1	3	1	3,079
Argentina - Chile	6	5	1	1	3	2,674
Brasil - Venezuela	2	6	2	1	0	2,641
Chile - Colômbia	2	2	0	0	2	2,325
Brasil - Colômbia	3	1	1	0	0	2,096
Cuba - México	2	7	0	1	2	1,963
Brasil - México	4	14	1	0	1	0,972

Tabela 4.11: Classificação das colaborações conforme o índice I_S no período 1994-2013.



(a) I_T (b) I_S

Figura 4.4: Distribuição das colaborações entre grupos segundo os índices I_T e I_S no período 1994-2013.

4.2 Análise das Redes de Colaboração

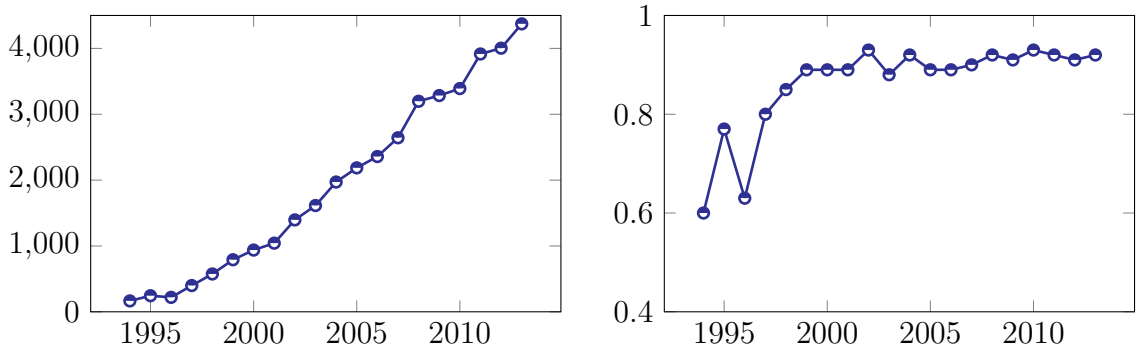
Nesta seção é apresentada a análise das redes de colaboração de cada país envolvido no estudo, enfatizado os dois grupos que apresentaram a produção de maior qualidade como descrito na Subsecção 4.1.3. A fim de analisar as cooperações entre os países e as colaborações entre os grupos de Ciência da Computação na América Latina, analisamos cada rede de colaboração em separado para obtermos uma ideia da estrutura e propriedades de cada rede. Com o intuito de caracterizar cada uma das redes, a Tabela 4.12 apresenta as estatísticas gerais da rede Latino-Americana em Ciência da Computação, denominada LACompNet, ao longo do período 1994-2013.

A LACompNet possui 18.523 vértices e 30.669 arestas. Na Tabela 4.12 podemos

Estatística	**Abreviatura**	**LACompNet**
Ordem	$\|V\|$	18.523
Tamanho	$\|E\|$	30.669
Grau médio	$< k >$	3,31
Coeficiente de agrupamento	C_a	0,27
Assortatividade	r	-0,38
Total de componentes conectados	T_{cc}	157
Tamanho do componente gigante	C_g	17.476
Diâmetro do componente gigante	d	21
Comprimento do caminho médio	l	5,82

Tabela 4.12: Estatísticas gerais da rede LACompNet no período 1994-2013.

observar que existem 157 componentes conectados, sendo um deles o componente gigante da rede, abrangendo $94,35\%$ dos pesquisadores, tendo o Brasil como o maior colaborador na região, além de ser o responsável por manter a rede conectada. A Figura 4.5a apresenta a evolução do componente gigante ao longo do tempo, assim como a Figura 4.5b apresenta a porcentagem de cobertura desse componente sobre a rede em função do tempo. Note-se que a partir de 1997 a rede apresenta uma cobertura acima de 80% dos vértices do grafo. Nesse aspecto, os pesquisadores das instituições brasileiras têm uma importante posição na rede dado que conectam diferentes grupos de pesquisa na região, permitindo a colaboração interinstitucional como mostrado nas Tabelas 4.7 e 4.8 durante as duas décadas.



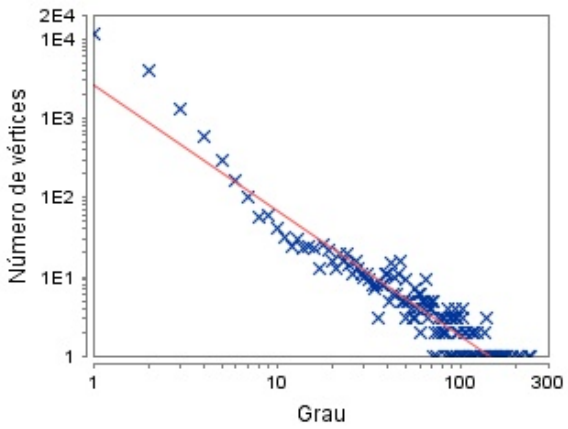
(a) Componente gigante C_g (b) % da cobertura do componente gigante

Figura 4.5: Evolução de maior componente conectado no período 1994-2013.

Por outro lado, a rede apresenta uma assortatividade negativa, dado que existe uma tendência de pesquisadores com grau maior a se conectarem com pesquisadores

de grau menor, ou seja, pesquisadores que estão iniciando a vida acadêmica tendem a se conectar com pesquisadores com ampla experiência e trajetória na área ao longo do tempo. A importância deste tipo de relacionamento na rede consiste em que os novos pesquisadores trazem para a rede novos tópicos de pesquisa que ajudam o enriquecimento dos grupos de pesquisa e a geração de novos conhecimentos, assim como a consolidação de novos grupos de pesquisa.

De forma geral, o número médio de colaboradores por pesquisador ($< k >$) é de 3,31, muito similar ao número médio de autores por publicação, mostrando uma preferência dos pesquisadores a se conectarem e trabalharem em conjunto. Uma outra caraterística da rede LACompNet é a distribuição dos seus graus (ver Figura 4.6) que segue a lei de potência, mostrando que novos pesquisadores tendem a se conectar com vértices de maior grau. Da mesma forma, a probabilidade de escolher aleatoriamente

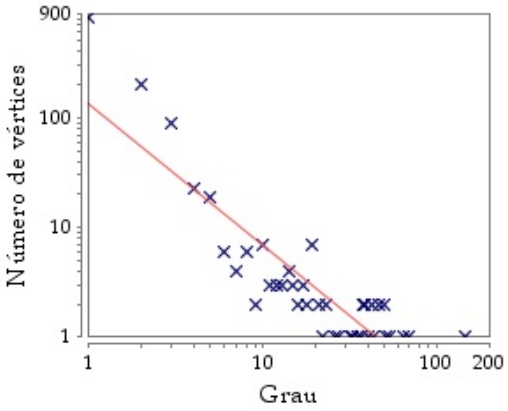


(a) LACompNet. $\gamma = -1,58$; $R^2 = 0,91$

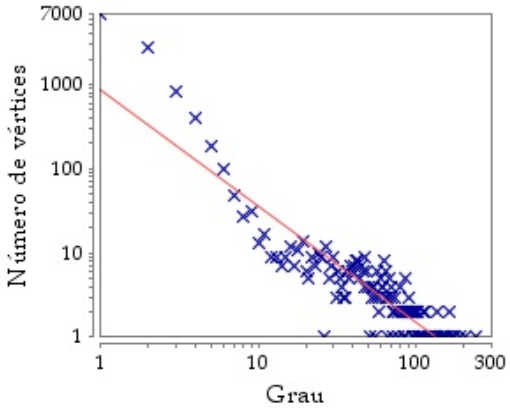
Figura 4.6: Distribuição de graus da rede LACompNet no período 1994-2013.

um vértice com menor grau é alta neste tipo de rede de colaboração científica. Ou seja, há poucos pesquisadores com alta conectividade e muitos pesquisadores com baixa conectividade, de modo que pesquisadores com maior grau têm um grande número de colaboradores, como nos casos de Wagner Meira Jr., Ricardo A. Baeza-Yates e Carlos José Pereira de Lucena que contam com mais de 200 colaboradores ao longo do período 1994-2013, já que trabalham em grupos com grande reconhecimento na América Latina, tornando a difusão do conhecimento mais rápida. Especificamente, as redes de coautoria da Argentina, Brasil e Chile, mostradas na Figura 4.7, seguem a lei de potência [Barabási & Albert, 1999; Newman, 2001b; Strogatz, 2001], onde a maioria dos vértices tem um pequeno número de vizinhos (grau baixo), mas há alguns vértices com uma alta conectividade (grau alto), superior à média, denominados *hubs* ou vértices

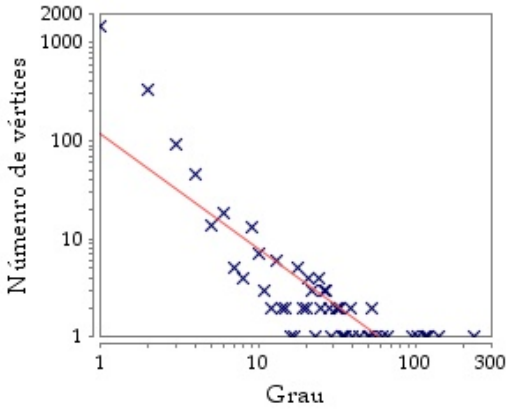
altamente conectados que seguem exatamente o modelo livre de escala [Barabási et al., 2009] que rege muitas redes presentes no mundo real.



(a) Argentina. $\gamma = -1,29$; $R^2 = 0,78$



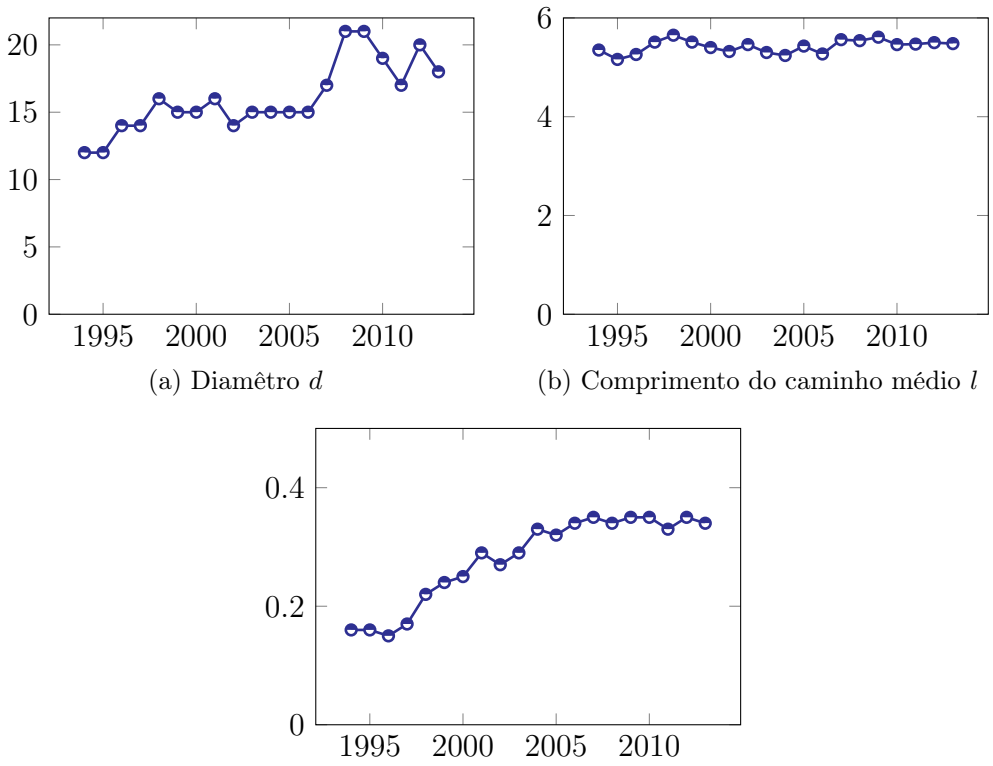
(b) Brasil. $\gamma = -1,38$; $R^2 = 0,85$



(c) Chile. $\gamma = -1,17$; $R^2 = 0,72$

Figura 4.7: Distribuição de graus no período 1994-2013.

O diâmetro da LACompNet é 21, que representa a distância máxima entre dois vértices no maior componente conectado. Em contraste, o valor do comprimento do caminho médio, se escolhermos dois vértices de forma aleatória para se conectarem no maior componente conectado, é $5,82$, mostrando claramente um comportamento de rede de mundo pequeno [Watts, 1999b, 2004]. A Figura 4.8 ilustra a evolução temporal do diâmetro, o comprimento do caminho médio e o coeficiente de agrupamento ao longo do tempo. Note-se que o diâmetro da rede tem uma tendência a aumentar ano a ano, evidenciando que os novos pesquisadores que entraram na rede foram conectados ao maior componente conectado através dos pesquisadores de maior conectividade na rede.



(a) Diamêtro d

(b) Comprimento do caminho médio l

(c) Coeficiente de agrupamento C_a

Figura 4.8: Diâmetro, Comprimento do caminho médio e Coeficiente de agrupamento na rede LACompNet em função do tempo.

A Figura 4.8c ilustra o crescimento do coeficiente de agrupamento em função do tempo. Note-se que ao longo do tempo os pesquisadores se conectarem e estabeleceram colaborações entre eles formando pequenos agrupamentos, com uma probabilidade de 27% de que dois coautores de um mesmo pesquisador estejam conectados entre si.

A Tabela 4.13 apresenta um resumo das propriedades da rede de cada um dos países envolvidos neste estudo. Nesse cenário pode-se observar que há uma tendência clara dos novos pesquisadores de se conectarem com pesquisadores de maior experiência, confirmado pelo coeficiente de assortatividade (r) que é negativo para todos os países. Em relação ao coeficiente de agrupamento, ressalte-se que há três países com valores muito baixos, que são Colômbia (0,09), Costa Rica (0,04) e Peru (0,05), o que indica que seus pesquisadores estão isolados e trabalham em grupos com poucas colaborações dentro e fora de suas instituições, o que é confirmado pela baixa cobertura do maior componente conectado das respectivas redes que não alcança 50% em cada caso. Por outro lado, países como Argentina, Brasil, Chile e México alcançam coberturas do componente conectado superiores a 70% da rede, evidenciando redes

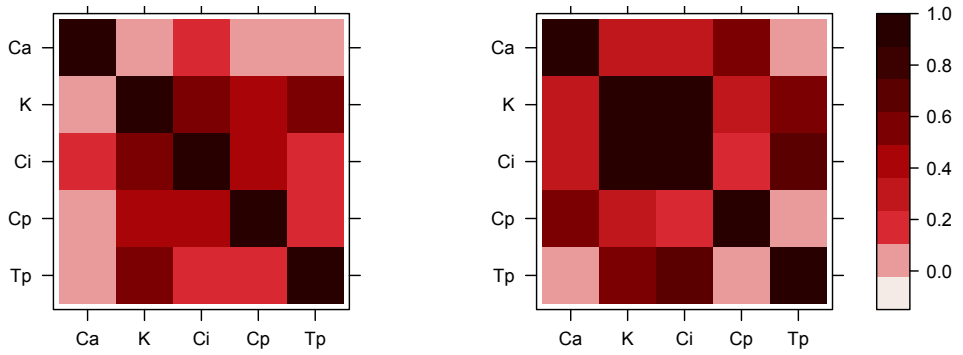
País	$\|V\|$	$\|E\|$	$< k >$	C_a	r	C_g	% Cob. C_g	d	l
Argentina	1.245	1.774	2,85	0,25	-0,36	920	73,90	14	5,81
Brasil	11.920	20.983	3,52	0,31	-0.46	11.746	98,54	10	4,98
Chile	2.105	2.917	2,77	0,21	-0,32	1.985	94,30	16	5,14
Colômbia	653	693	2,12	0,09	-0,46	296	45,33	10	4,82
Costa Rica	61	52	1,70	0,04	-0,53	16	26,23	4	1,96
Cuba	214	287	2,68	0,22	-0,30	53	24,77	7	2,54
México	2.039	2.362	2,32	0,12	-0,33	1.720	84,36	18	6,44
Paraguai	105	118	2,25	0,13	-0,58	85	80,95	4	2,52
Peru	129	128	1,98	0,05	-0.68	34	26,36	5	2,42
Uruguai	276	362	2,62	0,16	-0,40	218	78,99	11	4,98
Venezuela	516	635	2,46	0,22	-0.29	247	47,87	11	4,10

Tabela 4.13: Propriedades das redes de coautoria dos países no período 1994-2013.

muito compactas e que promovem o estabelecimento de colaborações entre grupos de pesquisa para fortalecer sua produção cientifica. De fato, esses países foram os que apresentaram a maior taxa média geral de publicações no período 1994-2013.

A Figura 4.9 apresenta o coeficiente de correlação de *Pearson* entre as métricas de coeficiente de agrupamento, centralidade de grau, centralidade de proximidade, centralidade de intermediação e número de publicações por pesquisador (normalizado pelo número de coautores) nos dois períodos. Nessa figura podemos observar o incremento do coeficiente de correlação para a maioria das métricas. Como se pode ver, existe uma correlação entre a centralidade de grau e o número de publicações, embora não seja tão alta ($0,58$ e $0,52$ no primeiro e segundo períodos respectivamente). Assim mesmo, nas duas décadas podemos evidenciar que a centralidade de grau está altamente correlacionada com a centralidade de intermediação. Além disso, o coeficiente tende a aumentar de um período para outro, dado que vértices com maior grau atuam como "*hubs*" dentro da rede, atraindo mais pesquisadores de diferentes áreas para se conectarem. Além disso, se evidencia que pesquisadores com maior número de colaboradores tendem a ter uma taxa de produtividade maior ao longo do tempo. Por outro lado, nota-se que o coeficiente de agrupamento na segunda década começa a se correlacionar com a métrica centralidade de proximidade, dado que os pesquisadores procuram expandir suas redes localmente através de novas colaborações, o que produz uma maior probabilidade de conexão entre eles.

As Tabelas 4.14, 4.15 e 4.16 identificam os pesquisadores da rede que possuem uma



(a) 1994-2003 (b) 2004-2013

Figura 4.9: Teste de correlação de *Pearson* para as métricas de coeficiente de agrupamento (Ca), centralidade de intermediação (Ci), centralidade de proximidade (Cp), centralidade de grau (K) e total de publicações (Tp).

posição privilegiada em sua estrutura segundo as métricas centralidade de grau, centralidade de proximidade e centralidade de intermediação. Ressalte-se que de acordo com essas três métricas há uma grande porcentagem de pesquisadores brasileiros, com algumas participações de pesquisadores chilenos e mexicanos.

Pesquisador	Instituição	k
Wagner Meira Jr.	UFMG	0,01312
Ricardo A. Baeza-Yates	UCHILE	0,01274
Carlos José Pereira de Lucena	PUC-RIO	0,01166
Marcos André Gonçalves	UFMG	0,01107
Luigi Carro	UFRGS	0,01042
Alessandro Garcia	PUC-RIO	0,00972
Jano Moreira de Souza	UFRJ-COPPE	0,00945
Jussara M. Almeida	UFMG	0,00929
Silvio Romero de Lemos Meira	UFPE	0,00918
Carlos A. Coello Coello	CINVESTAV	0,00896
Virgilio Almeida	UFMG	0,00875
Antonio A. F. Loureiro	UFMG	0,00869
Jorge Urrutia	UNAM	0,00848
Julio Cesar Sampaio do Prado Leite	PUC-RIO	0,00831
Jaelson Castro	UFPE	0,00815
Thaís Vasconcelos Batista	UFRN	0,00794
Gustavo Rossi	UNLP	0,00783
José Carlos Maldonado	ICMC-USP	0,00772
Lisandro Zambenedetti Granville	UFRGS	0,00756
Ricardo Augusto da Luz Reis	UFRGS	0,00756
Jesús Favela	CICESE	0,00756
Gonzalo Navarro	UCHILE	0,00740
Paulo Romero Martins Maciel	UFPE	0,00740
Alexandre X. Falcão	UNICAMP	0,00734
Marta Mattoso	UFRJ-COPPE	0,00713
André Carlos Ponce Leon Ferreira de Carvalho	ICMC-USP	0,00707
Marco A. Casanova	PUC-RIO	0,00696
Caetano Traina Jr.	ICMC-USP	0,00669
Philippe Olivier Alexandre Navaux	UFRGS	0,00669
Miguel Nussbaum	PUC-CHILE	0,00653

Tabela 4.14: Relação dos 30 principais pesquisadores segundo a métrica centralidade de grau no período 1994-2013.

Pesquisador	Instituição	C_P
Carlos José Pereira de Lucena	PUC-RIO	0,25053
Alberto H. F. Laender	UFMG	0,24638
Ricardo A. Baeza-Yates	UCHILE	0,24468
Thaís Vasconcelos Batista	UFRN	0,24371
Jussara M. Almeida	UFMG	0,24252
Nivio Ziviani	UFMG	0,24189
José Carlos Maldonado	ICMC-USP	0,24185
Simone Diniz Junqueira Barbosa	PUC-RIO	0,24182
Marcos André Gonçalves	UFMG	0,23807
Wagner Meira Jr.	UFMG	0,23771
Alessandro Garcia	PUC-RIO	0,23616
Edleno Silva de Moura	UFAM	0,23468
Marco A. Casanova	PUC-RIO	0,23438
Virgilio Almeida	UFMG	0,23420
Julio Cesar Sampaio do Prado Leite	PUC-RIO	0,23288
Edmundo R. M. Madeira	UNICAMP	0,23264
Sérgio Soares	UFPE	0,23216
Eduardo Figueiredo	UFMG	0,23172
Ana Carolina Salgado	UFPE	0,23136
Ricardo da Silva Torres	UNICAMP	0,23044
Fernando Castor	UFPE	0,22956
Marco Tulio Valente	UFMG	0,22952
Paulo Cesar Masiero	ICMC-USP	0,22888
Jaelson Castro	UFPE	0,22875
Antonio A. F. Loureiro	UFMG	0,22807
Lyrene Fernandes da Silva	UFRN	0,22800
Uirá Kulesza	UFRN	0,22763
Gisele L. Pappa	UFMG	0,22750
Rosana T. V. Braga	ICMC-USP	0,22733
Altigran Soares da Silva	UFAM	0,22694

Tabela 4.15: Relação dos 30 principais pesquisadores segundo a métrica centralidade de proximidade (*closeness*) no período 1994-2013.

Pesquisador	Instituição	C_I
Ricardo A. Baeza-Yates	UCHILE	0,09685
Carlos José Pereira de Lucena	PUC-RIO	0,05048
José Carlos Maldonado	ICMC-USP	0,03272
Wagner Meira Jr.	UFMG	0,02839
Jayme Luiz Szwarcfiter	UFRJ-COPPE	0,02680
Jussara M. Almeida	UFMG	0,02678
Simone Diniz Junqueira Barbosa	PUC-RIO	0,02665
Carlos A. Coello Coello	CINVESTAV	0,02655
Gonzalo Navarro	UCHILE	0,02617
Thaís Vasconcelos Batista	UFRN	0,02480
Sergio F. Ochoa	UCHILE	0,02476
Jesús Favela	CICESE	0,02456
Jorge Urrutia	UNAM	0,02449
Antonio A. F. Loureiro	UFMG	0,02345
Ana Carolina Salgado	UFPE	0,02304
Luciana Porcher Nedel	UFRGS	0,02230
José A. Pino	UCHILE	0,02222
André Carlos Ponce Leon Ferreira de Carvalho	ICMC-USP	0,02174
Gustavo Rossi	UNLP	0,02161
Alberto H. F. Laender	UFMG	0,02148
Marcos André Gonçalves	UFMG	0,02088
Edmundo R. M. Madeira	UNICAMP	0,02081
Edgar Chávez	CICESE	0,02077
Mauricio Marín	USACH	0,02051
Silvio Romero de Lemos Meira	UFPE	0,02033
Nivio Ziviani	UFMG	0,01999
Miguel Nussbaum	PUC-CHILE	0,01967
Philippe Olivier Alexandre Navaux	UFRGS	0,01863
Rafael Prikladnicki	PUC RS	0,01823
Luciana S. Buriol	UFRGS	0,01809

Tabela 4.16: Relação dos 30 principais pesquisadores segundo a métrica centralidade de intermediação (*betweenness*) no período 1994-2013.

Capítulo 5

Visualizações das Redes

Este capítulo apresenta alguns exemplos de visualizações que ilustram as redes de coautoria entre países e instituições, bem como de gráficos que mostram estatísticas das publicações dos grupos de pesquisa, incluindo análises baseadas na classificação Qualis dos respectivos veículos e cujos resultados podem ser explorados mais profundamente através da plataforma desenvolvida.

5.1 Introdução

A área de visualização de dados tem crescido amplamente nos últimos anos com a criação de novas ferramentas e técnicas que permitem a exploração de grandes volumes de informação de uma forma flexível, além de possibilitar em muitos casos a geração de conhecimento e a extração de informação mais rapidamente sem a necessidade de aplicação de algoritmos de mineração de dados. De fato, novas ferramentas têm sido desenvolvidas fornecendo uma excelente abordagem para explorar os dados dinamicamente, além de serem essenciais para a apresentação de resultados. Uma das principais características da área de visualização de dados é que permite a fácil percepção visual de padrões frequentes através de gráficos simples que permitem análises multivariadas, análises temporais, agrupamentos de dados, etc., oferecendo apoio visual a tarefas de análise em dados de larga escala e complexos. Segundo Keim [2002], para que o processo de mineração de dados seja eficaz, é importante incluir o ser humano na etapa de exploração de dados, de modo a combinar a sua flexibilidade e criatividade, com a enorme capacidade de armazenamento e poder computacional de hoje em dia. Portanto, a exploração visual dos dados permite ao ser humano ter uma visão dos mesmos, obter conclusões e interagir diretamente com os dados.

5.2 Rede LACompNet

A Figura 5.1 ilustra a rede LACompNet representada como um grafo não dirigido, onde o tamanho dos vértices simboliza o total de publicações de cada pesquisador e as arestas representam uma coautoria entre pesquisadores. Nessa rede claramente podemos observar a predominância dos pesquisadores brasileiros e também de argentinos, chilenos e mexicanos.

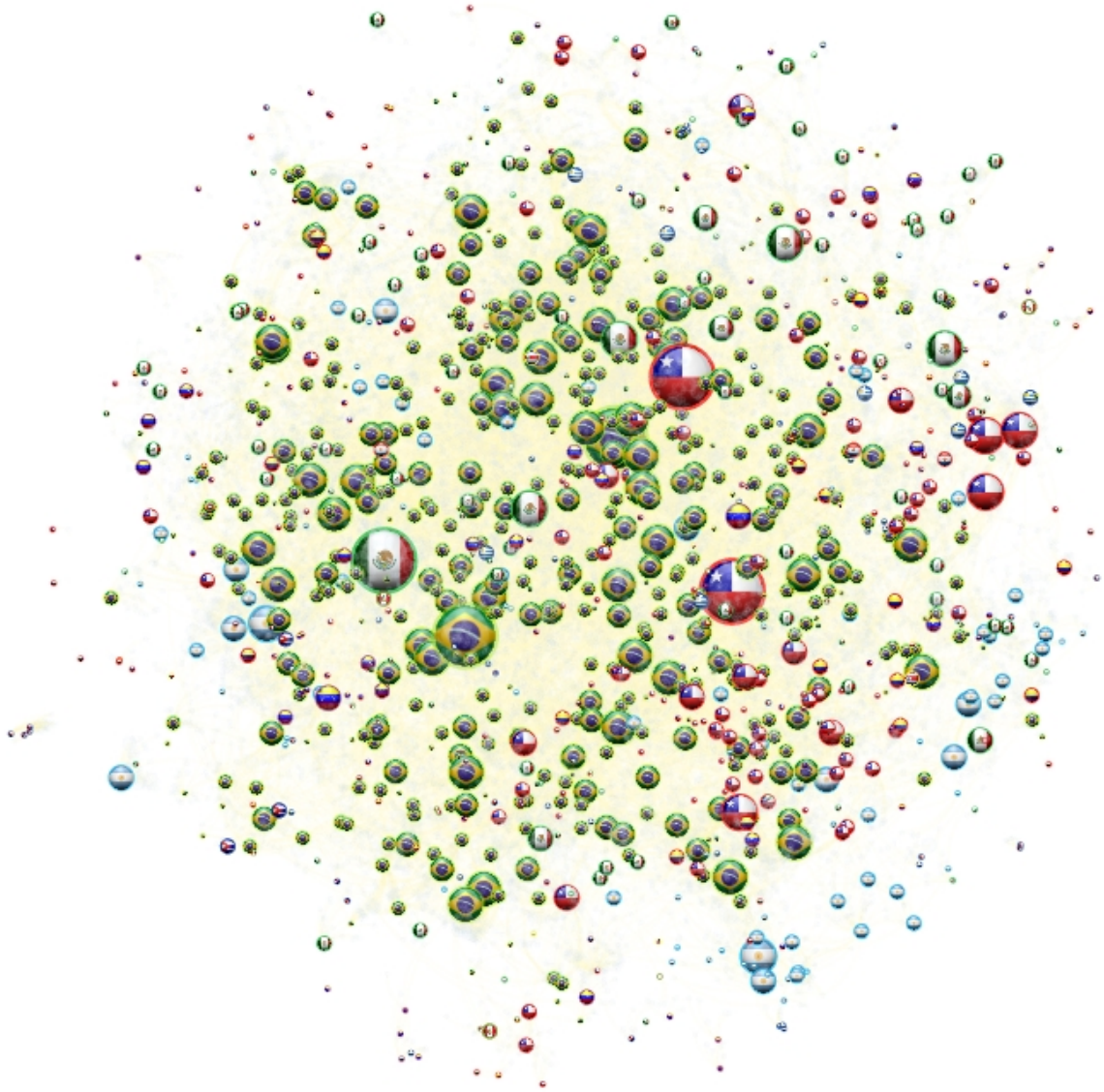
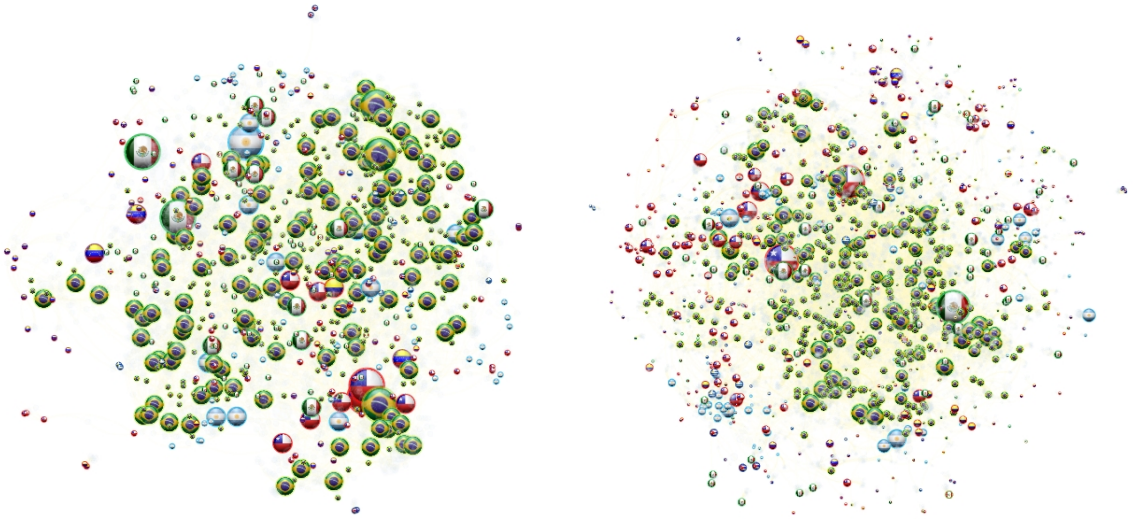


Figura 5.1: Rede Latino-Americana em Ciência da Computação no período 1994-2013.

As Figuras 5.2a e 5.2b ilustram a evolução da rede LACompNet em função do tempo, na qual pode-se observar a importância dos pesquisadores brasileiros ao manterem a

rede conectada nos dois períodos. Nota-se também, que na segunda década, o grafo apresenta três vértices de maior tamanho, que representam os pesquisadores Gonzalo Navarro e Ricardo A. Baeza-Yates da UCHILE e Carlos A. Coello Coello do CINVESTAV com uma produção acima de 200 publicações naquele período.



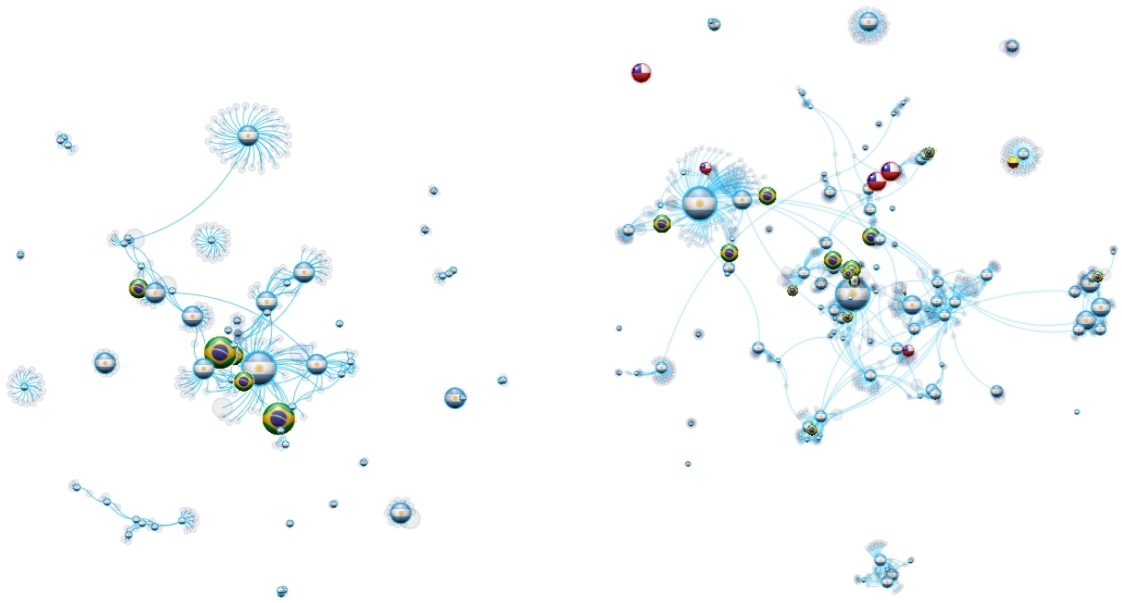
(a) C_a:0,24, C_g:4.294 (93, 39%), d:16, l:5,49 (b) C_a:0,29, C_g:15.163 (94, 24%), d:20, l:5,84

Figura 5.2: Rede LACompNet nas décadas (a) 1994-2003 e (b) 2004-2013.

5.3 Redes de Coautoria dos Países

Nesta seção são apresentadas as redes de coautoria dos quatro países, Argentina, Brasil, Chile e México, que apresentaram maior produção científica (Figuras 5.3 a 5.6). Essas figuras ressaltam que na primeira década foram formados pequenos agrupamentos entre pesquisadores do mesmo país e estabelecidas algumas colaborações com pesquisadores de outros países. Já na segunda década, pode-se observar que esses grupos foram consolidando-se, incorporando novos pesquisadores como evidenciado pelos valores das métricas do componente gigante que tiveram um aumento significativo em cada uma das redes. Além disso, o coeficiente de agrupamento cresceu em cada uma das redes, aumentado a probabilidade de maior colaboração entre os pesquisadores desses grupos. Por outro lado, para o cômputo de algumas das métricas mostradas em cada uma das redes, foi necessário determinar o componente gigante dado que a maioria das redes é constituída de grafos desconexos [Liu et al., 2005].

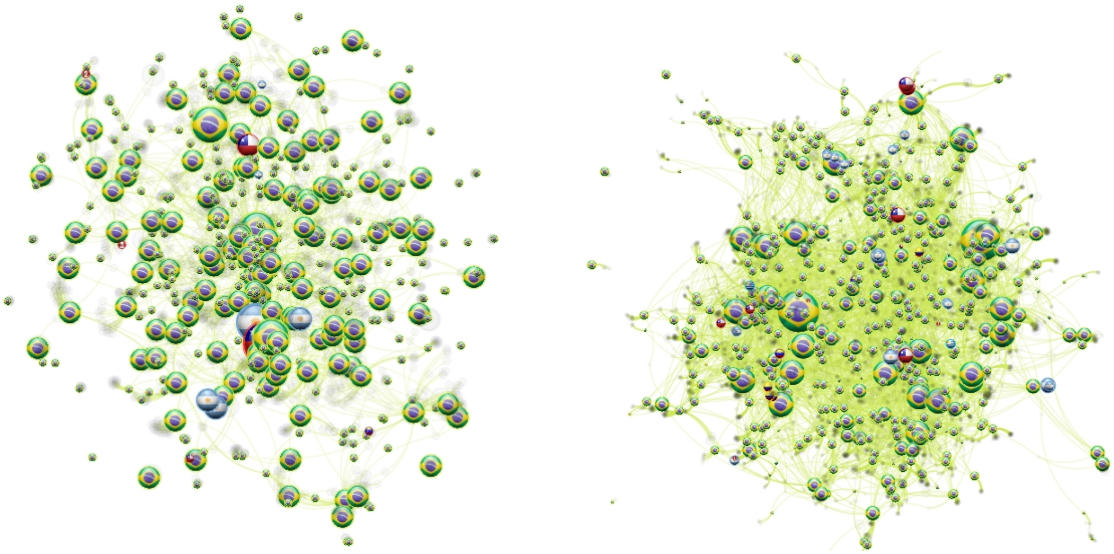
A Figura 5.3 apresenta a rede de coautoria argentina nos dois períodos. Pode-se observar que na primeira década (Figura 5.3a) a metade dos pesquisadores está



(a) C_a:0,19, C_g:187 (53, 58%), d:9, l:4,32 (b) C_a:0,26, C_g:730 (70.12%), d:14, l:6,25

Figura 5.3: Rede de Coautoria da Argentina nos períodos (a) 1994-2003 e (b) 2004-2013.

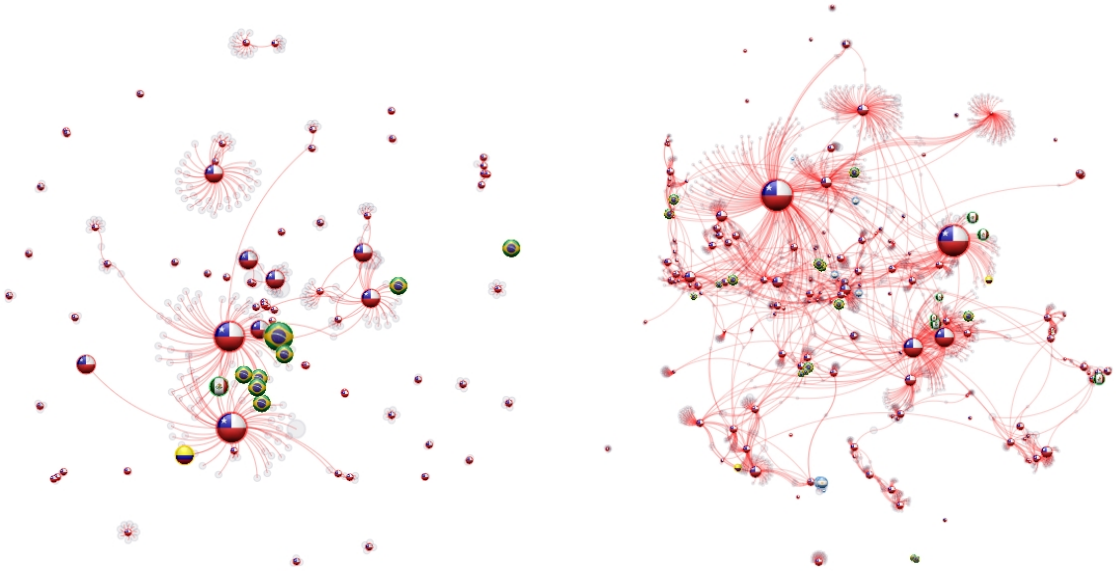
conectada ao componente gigante. Entretanto, na segunda década há uma quantidade maior de pesquisadores que se conectaram ao componente gigante, o que produz um aumento no coeficiente de agrupamento como evidenciado na Figura 5.3b. Quanto ao comprimento do caminho médio, pode-se observar o seu aumento em razão do aumento do diâmetro da rede de um período para o outro.



(a) C_a:0,23, C_g:2.705 (88, 57%), d:19, l:6,74 (b) C_a:0,33, C_g:10.106 (98, 33%), d:12, l:5,18

Figura 5.4: Rede de Coautoria do Brasil nos períodos (a) 1994-2003 e (b) 2004-2013.

A Figura 5.4 apresenta a rede de coautoria brasileira nos dois períodos. Observa-se que na segunda década (Figura 5.4b) há uma cobertura maior do componente gigante e uma tendência maior dos pesquisadores trabalharem em conjunto. Ressalta-se que neste caso o comprimento do caminho médio diminuiu da primeira década para a segunda, confirmado também pela redução do diâmetro da rede, mostrando uma maior densificação da rede brasileira na segunda década devido a novas colaborações entre grupos, confirmado pelo aumento do valor do coeficiente de agrupamento no período 2004-2013.

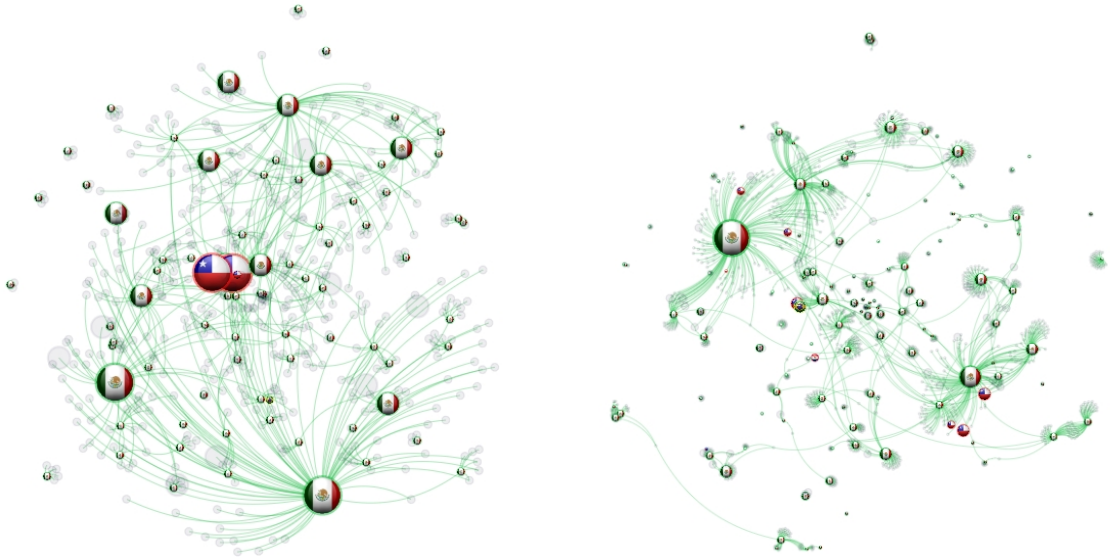


(a) C_a:0,16, C_g:177 (44, 81%), d:9, l:3,49 (b) C_a:0,21, C_g:1.666 (87, 50%), d:14, l:4,91

Figura 5.5: Rede de Coautoria do Chile nos períodos (a) 1994-2003 e (b) 2004-2013

A Figura 5.5 apresenta a rede de coautoria chilena nos dois períodos. Ressalta-se o aumento do diâmetro da rede da primeira década para a segunda, devido à presença de novos pesquisadores na rede que fortaleceram os grupos de pesquisa, aumentando também o comprimento do caminho médio. Nota-se também que houve um aumento considerável (841,24%) na quantidade de pesquisadores que passaram a fazer parte da rede na segunda década e que foram conectados ao componente gigante através de pesquisadores que atuam como vértices centrais, como os casos dos pesquisadores Gonzalo Navarro e Ricardo Baeza-Yates afiliados à UCHILE. Além disso, foram estabelecidas colaborações com pesquisadores de outras instituições.

A Figura 5.6 apresenta a rede de coautoria mexicana nos dois períodos, na qual nota-se uma maior cobertura do componente gigante na segunda década. Além disso,



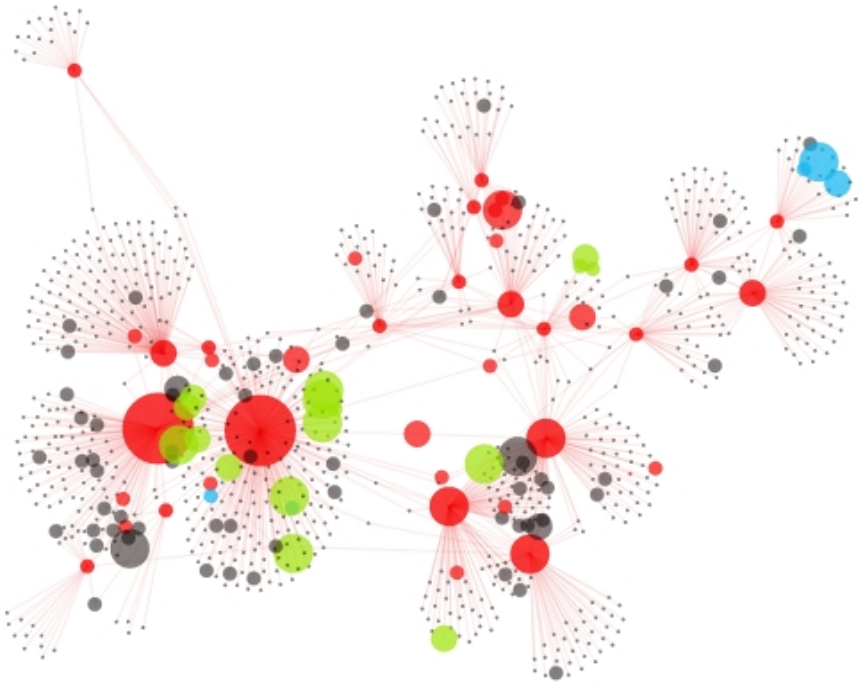
(a) C_a:0,04, C_g:143 (24,96%), d:8, l:3,13 (b) C_a:0,14, C_g:1.280 (74,99%), d:18, l:7,18

Figura 5.6: Rede de Coautoria do México nos períodos (a) 1994-2003 e (b) 2004-2013

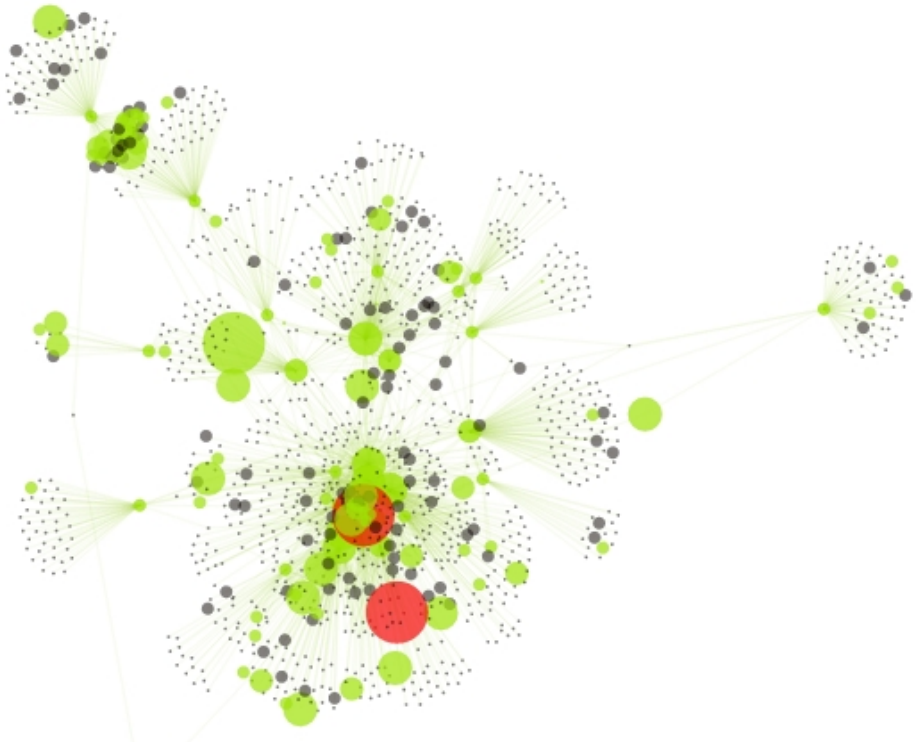
pode-se observar que o comprimento do caminho médio cresceu notavelmente, assim como o diâmetro da rede, evidenciando que na maioria dos casos as colaborações são feitas entre pequenos grupos das instituições mexicanas, confirmado pelo valor do coeficiente de agrupamento do segundo período.

5.4 Redes de Coautoria da UCHILE e UFMG

Nesta seção apresentamos as redes da UCHILE e UFMG (Figura 5.7), que são as instituições com as maiores taxas de publicação em seus respectivos países (ver Tabela 3.1). Ressalta-se que as redes dessas duas instituições cobrem 100% do total dos seus vértices, além de terem duas de suas propriedades de rede similares, no caso o diâmetro e o comprimento do caminho médio. Assim, pode-se observar que as duas redes apresentam um comportamento de rede de mundo pequeno, onde a distância máxima entre dois pares de vértices é 7 e o comprimento do caminho médio é menor do que 4.



(a) Propriedades da rede: C_{gg} :941 (100%), d_g:7, l:3,71, C_a :0,18



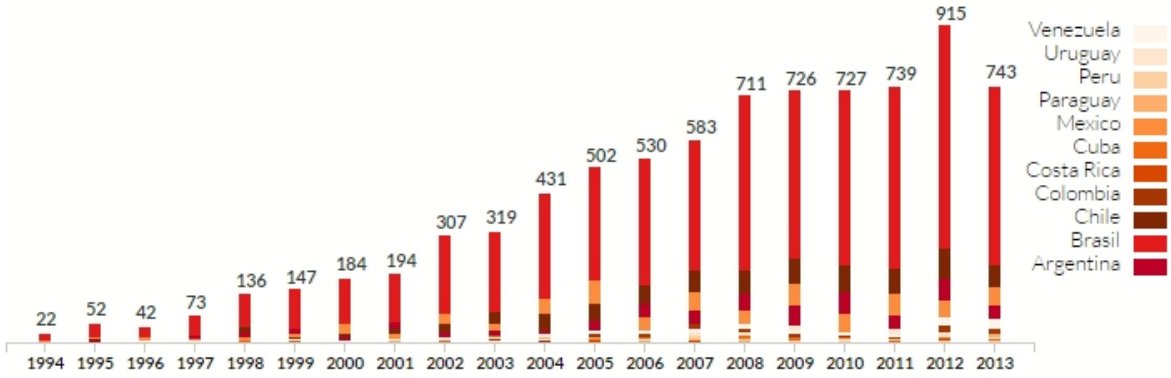
(b) Propriedades da rede: C_{gg} :1.567 (100%), d_g:7, l:3,68, C_a :0,33

Figura 5.7: Redes de Coautoria da (a) UCHILE e (b) UFMG no período 1994-2013.

5.5 Resumo da Produção Científica dos Países e das Colaborações entre Grupos segundo os Estratos Qualis

Nesta seção são apresentados os gráficos gerados pela plataforma LACompNet que resumem a produção científica dos países (Figura 5.8) e as colaborações entre os grupos (Figura 5.9) considerando a classificação dos veículos segundo os estratos Qualis. Como podemos observar, houve um incremento em todos os países, tanto quando consideramos os estratos superiores do Qualis (A1, A2 e B1) como quando consideramos os estratos inferiores (B2, B3, B4 e B5), com particular ênfase para o Brasil, Chile, Argentina e México. Além disso, os quantitativos mostram uma predominância das publicações nos estratos superiores do Qualis.

(a) Distribuição das publicações por país nos estratos A1, A2 e B1.

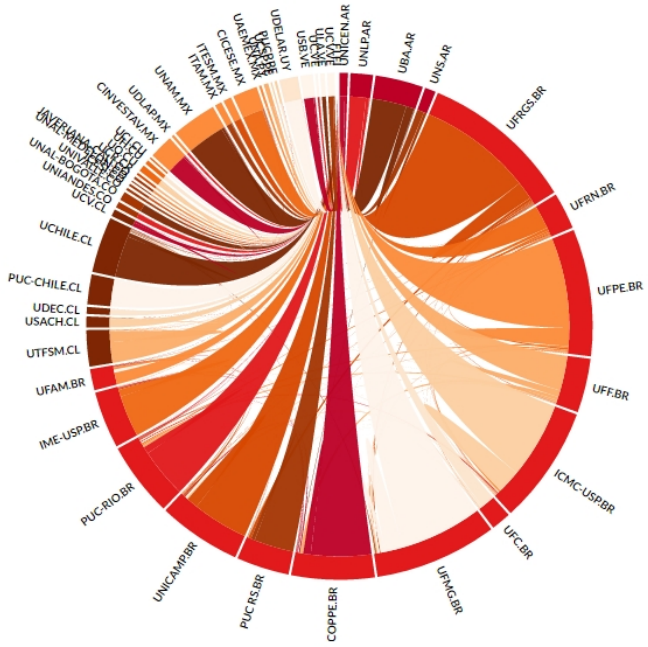


(b) Distribuição das publicações por país nos estratos B2, B3, B4 e B5.

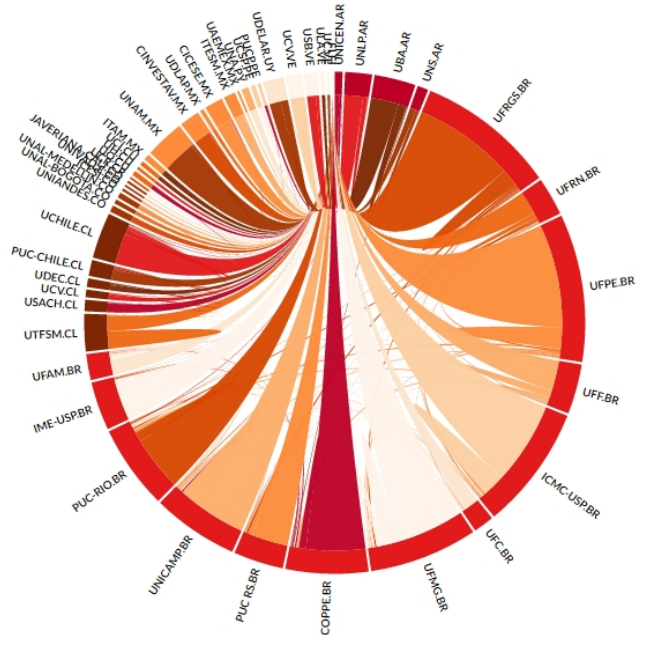
Figura 5.8: Distribuição da produção científica dos países nos estratos A1 a B5 no período 1994-2013.

Com relação aos gráficos da Figura 5.9 é possível ver que, de modo geral, as colaborações entre os grupos das diversas instituições inclui publicações nos dife-

rentes estratos, o que mostra uma certa diversificação da produção científica desses grupos. Uma visualização mais detalhada desses resultados pode ser obtida em `http://tortuga.lbd.dcc.ufmg.br/LACompNet`.



(a) Colaboração entre grupos nos estratos A1, A2 e B1.



(b) Colaboração entre grupos nos estratos B2, B3, B4 e B5.

Figura 5.9: Colaborações entre os grupos que publicaram nos estratos A1 a B5 no período 1994-2013.

Capítulo 6

Conclusões e Trabalhos Futuros

6.1 Revisão do Trabalho

Diferentes trabalhos analisaram a produção científica brasileira caracterizando e explorando as redes de coautoria em diferentes áreas do conhecimento, como também comparando com a produção de alguns países latino-americanos. Esses trabalhos mostraram que o Brasil é de longe o maior produtor de artigos científicos na América Latina [Van Noorden, 2014; Wainer et al., 2009]. Entretanto, neste cenário pouco se sabe sobre a estrutura e o perfil dos grupos de pesquisa em Ciência da Computação dos países latino-americanos.

Nesta dissertação foi apresentada uma análise extensa da produção científica de 48 instituições em Ciência da Computação da América Latina, cujos dados foram coletados da biblioteca digital da área da Ciência da Computação DBLP. Com base nesses dados foi feita uma análise bibliométrica a fim de determinar o perfil de cada grupo em termos de produção científica, além de conhecer a estrutura das redes de coautoria por meio de métricas e propriedades de redes complexas. Para isto foi desenvolvida uma plataforma denominada LACompNet para apoiar essas análises.

As análises mostraram que houve um aumento significativo no total de publicações na última década, além da consolidação de grupos de pesquisa na Argentina, Brasil, Chile e México, com um incremento na taxa média de publicação de $111,95\%$, $109,14\%$, $136,84\%$ e $113,89\%$, respectivamente, além da formação de novos grupos em países com menor tradição na área como Colômbia, Costa Rica, Cuba, Paraguai e Peru que tiveram também incrementos expressivos de $216,67\%$, 690%, $414,82\%$, $224,28\%$ e $357,14\%$, respectivamente.

Nossa análise permitiu observar que a colaboração entre grupos tem um efeito positivo sobre a produção científica da área, tanto quantitativa como qualitativamente, principalmente quando realizada entre grupos com ampla trajetória como mostrado na Subseção 4.1.2. Entretanto, nossa análise mostrou que de modo geral as colaborações se limitaram a apenas dois países. No primeiro período (1994-2003) essas colaborações envolveram os países de maior trajetória na área de Ciência da Computação (Argentina, Brasil, Chile e México), enquanto que no segundo período houve uma diversificação envolvendo também países de menor trajetória como Colômbia, Costa Rica, Cuba, Paraguai, Peru e Venezuela. Ressalta-se ainda que, em ambos os períodos, as colaborações mais prolíficas resultaram em publicações em estratos superiores do Qualis (A1 e B1, respectivamente).

Além disso, foram identificados os autores mais influentes na área, de acordo com três indicadores de redes complexas, a saber: centralidade de grau, centralidade de proximidade e centralidade de intermediação. Com relação às propriedades da rede LACompNet, evidenciou-se que existe uma correlação positiva quando um pesquisador corresponde a um vértice de maior grau, evidenciando um grande número de colaboradores e um maior número de publicações, como mostrado no Capítulo 4.

Finalmente, como parte desta dissertação, mostrou-se que a visualização de dados é um fator importante ao analisar os resultados, além de robustecer as análises e permitir a interação direta com os dados. Foi verificado, por exemplo, que países como o Brasil e Chile atuam como pontes para a transferência de informação entre grupos, como evidenciado pela Figura 5.1 que mostra claramente a importância dos pesquisadores do Brasil e do Chile na composição da rede. Essa mesma figura mostra que, apesar de o México ser um dos países mais produtivos da região, a sua produção está bastante concentrada em alguns pesquisadores.

6.2 Trabalhos Futuros

Apesar da extensa análise realizada sobre a produção científica dos grupos de pesquisa em Ciência da Computação latino-americanos, abrangendo tanto uma análise bibliométrica como uma análise das redes de colaboração, ainda existem oportunidades de trabalhos futuros que podem abordar aspectos importantes das redes como predição de novas colaborações entre grupos e perfil de novos pesquisadores influentes nas redes em

função do tempo. Ademais, seria importante avaliar a produção científica de um grupo levando em consideração que muitas vezes um pesquisador encontra-se afiliado a mais de uma instituição durante a sua trajetória acadêmica, o que serviria como um indicador de mobilidade e trataria com maior exatidão a produção de um determinado grupo.

Um outro ponto importante seria identificar os pesquisadores que não pertencem às instituições latino-americanas consideradas neste estudo, visando conhecer as colaborações com instituições externas e determinar o volume de colaborações com instituições de fora da América Latina.

Finalmente, o processo de análise adotado nesta dissertação poderia ser adaptado para avaliar outras áreas do conhecimento para saber como elas têm se desenvolvido ao longo dos anos, além de descobrir interações entre diferentes áreas e colaborações multidisciplinares que permitam estabelecer novas colaborações.

Referências Bibliográficas

Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47.

Albert, R.; Jeong, H. & Barabási, A.-L. (1999). Internet: Diameter of the World-Wide Web. *Nature*, 401(6749):130--131.

Barabási, A.-L. & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439):509--512.

Barabási, A.-L.; Albert, R. & Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1):173--187.

Barabási, A.-L. et al. (2009). Scale-Free Networks: A Decade and Beyond. *Science*, 325(5939):412.

Barabási, A. L.; Néda, Z.; Ravasz, E.; Schubert, A. & T, V. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590–614.

Bazzan, A. L. & Argenta, V. (2011). Network of collaboration among PC members of Brazilian computer science conferences. *Journal of the Brazilian Computer Society*, 17(2):133--139.

Benevenuto, F.; Almeida, J. & Silva, A. (2011). Coleta e Análise de Grandes Bases de Dados de Redes Sociais Online. *Jornadas de Atualizaçao em Informática*, pp. 11--57.

Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M. & Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424(4):175--308.

Coutinho, R. X.; Dávila, E. S.; dos Santos, W. M.; Rocha, J. B.; Souza, D. O.; Folmer, V. & Puntel, R. L. (2012). Brazilian scientific production in science education. *Scientometrics*, 92(3):697--710.

Cuadros-Vargas, E.; Silva-Sprock, A.; Delgado-Castillo, D.; Hernandez-Bieliukas, Y. & Collazos, C. (2013). Evolution of the Computing Curricula for Computer Science in Latin America 2013. In *Proceedings of the XXXIX Latin American Computing Conference*, pp. 1–10.

de Almeida, E. C. E. & Guimarães, J. A. (2013). Brazil's growing production of scientific articles-How are we doing with review articles and other qualitative indicators? *Scientometrics*, 97(2):287--315.

Delgado-Garcia, J. F.; Laender, A. H. F. & Meira Jr., W. (2014a). A Preliminary Analysis of the Scientific Production of Latin American Computer Science Research Groups. In *Proceedings of Alberto Mendelzon Workshop on Foundations of Data Management*, Cartagena de Indias, Colombia.

Delgado-Garcia, J. F.; Laender, A. H. F. & Meira Jr., W. (2014b). Analyzing the Coauthorship Networks of Latin American Computer Science Research Groups. In *Proceedings of Latin American Web Congress*, pp. 77--81, Ouro Preto, Brazil.

Digiampietri, L. A.; Mena-Chalco, J. P.; de Melo, P. O. V.; Malheiro, A. P.; Meira, D. N.; Franco, L. F. & Oliveira, L. B. (2014). BraX-Ray: An X-Ray of the Brazilian Computer Science Graduate Programs. *PloS one*, 9(4):e94541.

Easley, D. & Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.

Erdös, P. & Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae Debrecen*, 6:290--297.

Erdös, P. & Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17--61.

Faloutsos, M.; Faloutsos, P. & Faloutsos, C. (1999). On Power-Law Relationships of the Internet Topology. In *ACM SIGCOMM Computer Communication Review*, volume 29, pp. 251--262, Boston MA, USA.

Figueiredo, D. R. (2011). Introdução a Redes Complexas. In de Souza, A. F. & Meira Jr., W., editores, *Atualizações em Informática 2011*, capítulo 7, pp. 303--358. PUC-Rio.

Franceschet, M. (2011). Collaboration in computer science: A network science approach. *Journal of the American Society for Information Science and Technology*, 62(10):1992--2012.

Hermes-Lima, M.; Polcheira, C.; Trigueiro, M. & Beleboni, R. O. (2008). Perceptions of Latin American scientists about science and post-graduate education: Introduction to the 5th issue of CBP-Latin America. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, 151(3):263--271.

Huang, J.; Zhuang, Z.; Li, J. & Giles, C. L. (2008). Collaboration over Time: Characterizing and Modeling Network Evolution. In *Proceedings of Web Search and Data Mining*, pp. 107–116, Stanford, CA, USA.

Huang, T.-H. & Huang, M. L. (2006). Analysis and Visualization of Co-authorship Networks for Understanding Academic Collaboration and Knowledge Domain of Individual Researchers. In *Proceedings of Computer Graphics, Imaging and Visualisation, International Conference*, pp. 18--23, Sydney, Australia.

Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z. N. & Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804):651--654.

Jones, J. H. & Handcock, M. S. (2003). An assessment of preferential attachment as a mechanism for human sexual network formation. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1520):1123--1128.

Keim, D. A. (2002). Information Visualization and Visual Data Mining. *Proceedings of the IEEE Transactions on Visualization and Computer Graphics*, 8(1):1--8.

Kurosawa, T. & Takama, Y. (2012). Co-Authorship Networks Visualization System for Supporting Survey of Researchers' Future Activities. *Journal of Emerging Technologies in Web Intelligence*, 4(1):3--14.

Laender, A. H.; de Lucena, C. J.; Maldonado, J. C.; de Souza e Silva, E. & Ziviani, N. (2008). Assessing the Research and Education Quality of the Top Brazilian Computer Science Graduate Programs. *ACM SIGCSE Bulletin*, 40(2):135--145.

Laender, A. H.; Moro, M. M.; Gonçalves, M. A.; Davis Jr, C. A.; da Silva, A. S.; Silva, A. J.; Bigonha, C. A.; Dalip, D. H.; Barbosa, E. M.; Cortez, E. et al. (2011a). Building a Research Social Network from an Individual Perspective. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, pp. 427--428, Ottawa, Canada.

Laender, A. H. F.; Moro, M. M.; da Silva, A. S.; Jr., C. A. D.; Gonçalves, M. A.; Galante, R.; Silva, A. J. C.; Bigonha, C. A. S.; Dalip, D. H.; Barbosa, E. M.; Borges, E. N.; Cortez, E.; Jr., P. S. P.; de Alencar, R. O.; Cardoso, T. N. C. &

Salles, T. (2011b). CiênciaBrasil - The Brazilian Portal of Science and Technology. In *Anais do Seminário Integrado de Software e Hardware*, Natal, RN, Brazil.

Ley, M. (2009). DBLP: Some Lessons Learned. *Proceedings of the VLDB Endowment*, 2(2):1493--1500.

Liljeros, F.; Edling, C. R.; Amaral, L. A. N.; Stanley, H. E. & Åberg, Y. (2001). The web of human sexual contacts. *Nature*, 411(6840):907--908.

Liu, X.; Bollen, J.; Nelson, M. L. & Van de Sompel, H. (2005). Co-authorship Networks in the Digital Library Research Community. *Information Processing & Management*, 41(6):1462--1480.

Maia, G.; de Melo, P. O. S. V.; Guidoni, D. L.; Souza, F. S. H.; Silva, T. H.; Almeida, J. M. & Loureiro, A. A. F. (2013). On the Analysis of the Collaboration Network of the Brazilian Symposium on Computer Networks and Distributed Systems. *Journal of the Brazilian Computer Society*, 19(3):361–382.

McCool, J. I. (2003). Probability and Statistics With Reliability, Queuing and Computer Science Applications. *Technometrics*, 45(1):107--107.

Mena-Chalco, J. P.; Digiampietri, L. A.; Lopes, F. M. & Cesar, R. M. (2014). Brazilian Bibliometric Coauthorship Networks. *Journal of the Association for Information Science and Technology*, 66(7).

Menezes, G. V.; Ziviani, N.; Laender, A. H. & Almeida, V. (2009). A Geographical Analysis of Knowledge Production in Computer Science. In *Proceedings of the 18th International Conference on World Wide Web*, pp. 1041--1050, Madrid, Spain.

Milgram, S. (1967). The Small World Problem. *Psychology Today*, 2(1):60--67.

Nascimento, M. A.; Sander, J. & Pound, J. (2003). Analysis of SIGMOD's Co-authorship Graph. *SIGMOD Record*, 32(3):8–10.

Newman, M. & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69:026113.

Newman, M. E. (2001a). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64(1):016131.

Newman, M. E. (2001b). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404.

Newman, M. E. (2003a). Mixing patterns in networks. *Physical Review E*, 67(2):026126.

Newman, M. E. (2003b). The structure and function of complex networks. *SIAM review*, 45(2):167--256.

Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5200--5205.

Rodriguez, M. A. & Pepe, A. (2008). On the relationship between the structural and socioacademic communities of a coauthorship network. *Journal of Informetrics*, 2(3):195--201.

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410(6825):268--276.

Sun, J. & Tang, J. (2011). A Survey of Models and Algorithms for Social Influence Analysis. In Aggarwal, C. C., editor, *Social Network Data Analytics*, pp. 177–214. Springer US.

Van Noorden, R. (2014). The impact gap: South America by the numbers. *Nature*, 510(7504):202--203.

Wainer, J.; Xavier, E. C. & Bezerra, F. (2009). Scientific production in computer science: A comparative study of Brazil and other countries. *Scientometrics*, 81(2):535--547.

Wasserman, S. (1994). *Social Network Analysis: Methods and Applications*, volume 8. Cambridge University Press.

Watts, D. J. (1999a). Networks, Dynamics, and the Small-World Phenomenon. *American Journal of Sociology*, 105(2):493--527.

Watts, D. J. (1999b). *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton university press.

Watts, D. J. (2004). *Six Degrees: The Science of a Connected Age*. WW Norton & Company.

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of "small-world"networks. *Nature*, 393(6684):440--442.

Zaki, M. J. & Meira Jr, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Apêndice A

Estatísticas das Redes de Colaboração entre as Instituições

Instituição	k	Cp	Ci	l	C_a
UFRGS	19	0,568	0,264	1,762	0,374
UFRJ-COPPE	16	0,519	0,079	1,929	0,583
UCHILE	14	0,500	0,119	2,000	0,308
UFPE	14	0,494	0,025	2,024	0,648
PUC-RIO	14	0,472	0,014	2,119	0,703
ICMC-USP	14	0,462	0,064	2,167	0,637
UFMG	13	0,506	0,046	1,976	0,654
IME-USP	13	0,483	0,063	2,071	0,615
UNICAMP	13	0,457	0,012	2,190	0,756
UFF	11	0,467	0,012	2,143	0,691
PUC	11	0,442	0,001	2,262	0,873
UFC	11	0,442	0,001	2,262	0,927
UFAM	10	0,447	0,004	2,238	0,867
UTFSM	9	0,472	0,098	2,119	0,278
UBA	9	0,452	0,031	2,214	0,278
UFRN	9	0,438	0,020	2,286	0,667
CICESE	8	0,467	0,144	2,143	0,214
UNLP	7	0,447	0,056	2,238	0,286

continua na página seguinte

continuação da página anterior

Instituição	k	Cp	Ci	l	C_a
PUC-CHILE	7	0,416	0,041	2,405	0,238
USACH	5	0,433	0,012	2,310	0,400
UNIANDES	5	0,424	0,167	2,357	0,000
CINVESTAV	5	0,372	0,101	2,690	0,300
UNS	4	0,393	0,017	2,548	0,000
UDELAR	4	0,389	0,023	2,571	0,500
UCV	4	0,368	0,056	2,714	0,333
UCV	4	0,365	0,013	2,738	0,333
ITAM	4	0,353	0,010	2,833	0,500
UNAL-MEDELLIN	3	0,412	0,037	2,429	0,000
UNICEN	3	0,396	0,005	2,524	0,000
UNAM	3	0,336	0,000	2,976	1,000
USB	3	0,309	0,008	3,238	0,333
UNA	2	0,350	0,014	2,857	0,000
UDEC	2	0,350	0,000	2,857	1,000
PUCP	2	0,344	0,004	2,905	0,000
UCSP	2	0,326	0,000	3,071	1,000
ICESI	2	0,309	0,093	3,238	0,000
UCR	2	0,309	0,003	3,238	0,000
UO	2	0,276	0,048	3,619	0,000
UC	2	0,276	0,000	3,619	1,000
JAVERIANA-CALI	2	0,240	0,048	4,167	0,000
UNAL-BOGOTA	1	0,336	0,000	2,976	0,000
UH	1	0,218	0,000	4,595	0,000
UNIVALLE	1	0,194	0,000	5,143	0,000
ITESM	0	0,000	0,000	0,000	0,000
UAEMEX	0	0,000	0,000	0,000	0,000
UDLAP	0	0,000	0,000	0,000	0,000
UCI	0	0,000	0,000	0,000	0,000
ULA	0	0,000	0,000	0,000	0,000

Tabela A.1: Estatísticas da Rede de Colaboração entre Instituições no período 1994-2013.