Análise da Rede dos Doutores que Atuam em Computação no Brasil

Luciano A. Digiampietri[1], **Caio M. Alves**[1], **Caio C. Trucolo**[1], **Rômulo A.C. Oliveira**[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)
Av. Arlindo Bettio, 1000 – CEP 03828-000 – São Paulo – SP – Brasil

Abstract. *The Brazilian social and cultural diversity and the size of the country make any general analysis miss interesting details about the regional peculiarities. In this work, the Brazilian academic network of PhDs who work in computer science is analyzed, as well as, their sub-networks in order to identify their characteristics in terms of relationships between doctors and the keywords used in publications in each of the Brazilian states.*

Resumo. *O tamanho e a diversidade social e cultural do Brasil fazem com que qualquer análise global perca detalhes interessantes sobre as peculiaridades regionais. Neste trabalho a rede acadêmica brasileira dos doutores que atuam em ciência da computação é analisada bem como suas sub-redes estaduais de forma a se identificar suas caraterísticas em termos de relacionamentos entre doutores bem como os principais termos utilizados nas publicações em cada um dos estados brasileiros.*

1. Introdução

Atualmente a análise de redes sociais tem sido considera tão importante, ou mais importante, do que a análise de indicadores de um profissional ou grupo de profissionais, pois a maioria das tarefas complexas não é realizada por uma única pessoa (tanto em empresas quanto na academia) e boas relações entre os participantes no desenvolvimento destas tarefas é fundamental para garantir resultados satisfatórios.

Em empresas, a análise de redes sociais começou a ser amplamente utilizada principalmente para minimizar conflitos ou criar grupos cuja expertise coletiva otimize a resolução de uma dada demanda [Hsieh et al. 2013]. Na academia, os principais trabalhos sobre redes sociais estão ligados à caracterização de grupos de pesquisadores, análise de interações entre alunos e docentes, identificação de potenciais colaboradores e predição de coautorias [Melo-Minardi et al. 2013, Gao et al. 2012, Brandão et al. 2013, Xu et al. 2012, Hew 2011].

Além do uso de análise de redes sociais, outra estratégia que é bastante utilizada para avaliar grupos de pesquisadores são as análises bibliométricas. Nestas análises, diferentes tipos de produção (bibliográfica, técnica, cultural, etc) são analisadas de maneira quantitativa e qualitativa (por exemplo, pelo impacto das publicações) para se entender um dado grupo e/ou comparar grupos de uma dada área do conhecimento ou de diferentes regiões de um país [Digiampietri et al. 2014].

Outra estratégia de tentar caracterizar grupos de pesquisadores é o uso de mineração de texto sobre suas produções de forma a se identificar quais grupos estão

atuando em que áreas do conhecimento (ou sobre quais assuntos), bem como identificar às tendências de pesquisa dos grupos ao longo dos anos [Miyata et al. 2013].

Neste artigo estas estratégias são combinadas para analisar a rede formada pelos doutores que atuam no Brasil na área de Ciência da Computação. Esta análise considera a rede brasileira como um todo e também a rede formada em cada estado.

O restante deste artigo está organizado da seguinte maneira. A Seção 2 apresenta a metodologia utilizada. A Seção 3 contém a apresentação e a discussão dos resultados. Na Seção 4 são apresentados os principais trabalhos correlatos. Por fim, a Seção 5 contém as conclusões e as direções para os trabalhos futuros.

2. Metodologia

A metodologia deste trabalho foi dividida em cinco atividades:

Obtenção dos dados. Para este projeto foram obtidos os arquivos XML dos Currículos Lattes[1] de 3,2 milhões de pesquisados em julho de 2013. Optou-se por obter todos os currículos disponíveis na plataforma e depois se realizar a seleção da amostra ao invés de tentar se obter apenas os currículos de interesse[2].

Seleção da amostra. Para o estudo realizado neste artigo foram identificados os doutores que atuam no Brasil na área de Ciência da Computação. Para isto, foram selecionados os currículos que atendessem as seguintes restrições: (i) currículos com doutorados concluídos; (ii) com endereço profissional no Brasil; e (iii) com Ciência da Computação indicada como área de atuação. 6.358 currículos satisfizeram estas três restrições.

Seleção das informações de interesse. Foram consideradas informações de interesse dos currículos: endereço profissional; título dos artigos publicados em periódicos e em anais de eventos; ligações explícitas entre os currículos (cada currículo pode ter uma referência a outros currículos: coautores; coparticipantes de projetos; orientadores; orientandos e coparticipantes em bancas e comissões julgadoras). Todas as ligações explícitas foram consideradas para a montagem da rede social.

Cálculo de métricas. Métricas de redes sociais/grafos e métricas relacionadas à mineração de textos foram calculadas considerando-se 28 grupos: a rede brasileira, as 26 redes estaduais e a rede do Distrito Federal. Para as métricas de redes, dos 6.358 nós (correspondendo aos doutores da amostra) apenas 4.427 possuíam ao menos uma ligação (aresta) com outro nó, assim, as métricas foram calculadas considerando-se apenas estes nós. A Tabela 1 apresenta o nome e uma breve descrição de cada métrica utilizada. Adicionalmente a quantidade de arestas entre nós de um mesmo estado e entre nós de estados diferentes foi computada. A mineração de textos foi utilizada para se verificar quais as expressões mais utilizadas nas publicações feitas pelos pesquisadores em cada estado e na rede como um todo.

Análise dos resultados. As redes construídas foram analisadas de maneira comparativa, conforme será apresentado na próxima seção.

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

[2] Este artigo está contextualizado em um projeto que visa a caracterização de toda produção científica nacional. Os dados foram organizados conforme metodologia apresentada em [Digiampietri et al. 2012a, Digiampietri et al. 2012b]

Tabela 1. Métricas de Rede Utilizadas

Métrica	Descrição
Nós Totais	Número de nós presente na rede atual.
Porcentagem dos Nós	Porcentagem dos nós da rede em relação aos nós totais.
Arestas	Número de arestas presente na rede atual.
Nós com ligações	Número de nós da rede atual que possuem ao menos uma ligação com outro nó na rede nacional.
Nós no Componente Gigante	Número de nós no componente gigante (maior componete conexo).
Porcentagem de Nós no Componente Gigante	Porcentagem dos nós no componente gigante em relação a todos os nós da rede atual.
Densidade	Número de arestas do grafo atual em relação ao número máximo possível.
Grau Médio	Grau médio dos nós da rede atual.
Coeficiente de Clusterização	Quantidade de cliques de tamanho três dividida pela quantidade de três nós conectados.
Assortatividade de Grau	Métrica que calcula a tendência de nós de se conectarem a nós de mesmo grau (1 indica que todos os nós se conectam apenas a nós de mesmo grau e -1 indica que todos os nós se conectam a nós de grau diferentes dos seus).
Centralização de Grau	Métrica derivada da centralidade de grau que mede o quão central o nó mais central é em relação a todos os outros nós da rede, baseada no grau dos nós.
Centralização de Proximidade	Métrica derivada da centralidade de proximidade (*closeness*) que mede o quão central o nó mais central é em relação a todos os outros nós da rede, baseada na distância existente entre todos os pares de nós.
Diâmetro	Diâmetro da rede (grafo) atual.
Tamanho da Clique Máxima	Tamanho da maior clique do grafo atual, isto é, tamanho do subconjunto máximo de nós no qual todos os elementos do conjunto estão ligados uns aos outros.
Média dos Caminhos Mínimos	Média dos caminhos mínimos entre todos os pares de nós no componente gigante.

3. Análise dos Resultados

A Figura 1 apresenta a rede de coautorias considerando-se apenas um ponto por estado, onde cada nó tem tamanho proporcional a quantidade de doutores que atuam em computação no respectivo estado. É possível observar que não existem arestas ligando dois estados da Região Norte, mas vale destacar que há poucos doutores (que atenderam aos critérios de seleção utilizados neste artigo) atuando nesta região (conforme pode ser observado na Tabela 2).

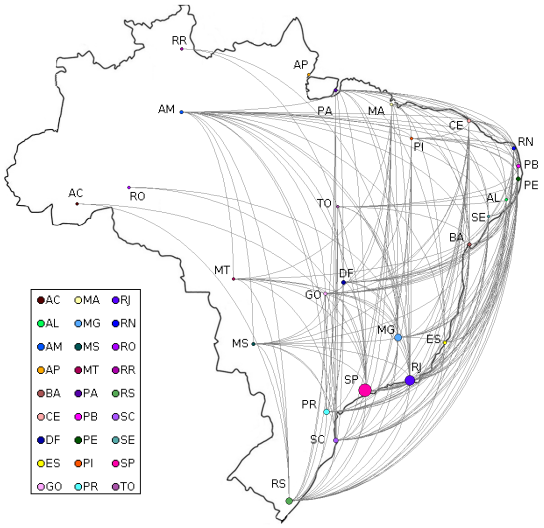

Figura 1. Rede Social dos Doutores em Computação - Estados

Os grafos da Figura 2 têm como nós cada uma das cidades que possuem ao menos um doutor atuando em computação. Na imagem da esquerda são apresentadas apenas as arestas que ligam duas cidades do mesmo estado, indicando que há doutores nestas cidades que se relacionam. Na imagem da esquerda é possível observar uma maior conexão entre as cidades das regiões Sul e Sudeste. Já na imagem da direita, as arestas em cinza ligam duas cidades de diferentes estados. Destaca-se o grande número de arestas entre as cidades da região Nordeste com as cidades da região Sudeste.

Nos grafos da Figura 3, cada nó representa um doutor e foi posicionado em torno

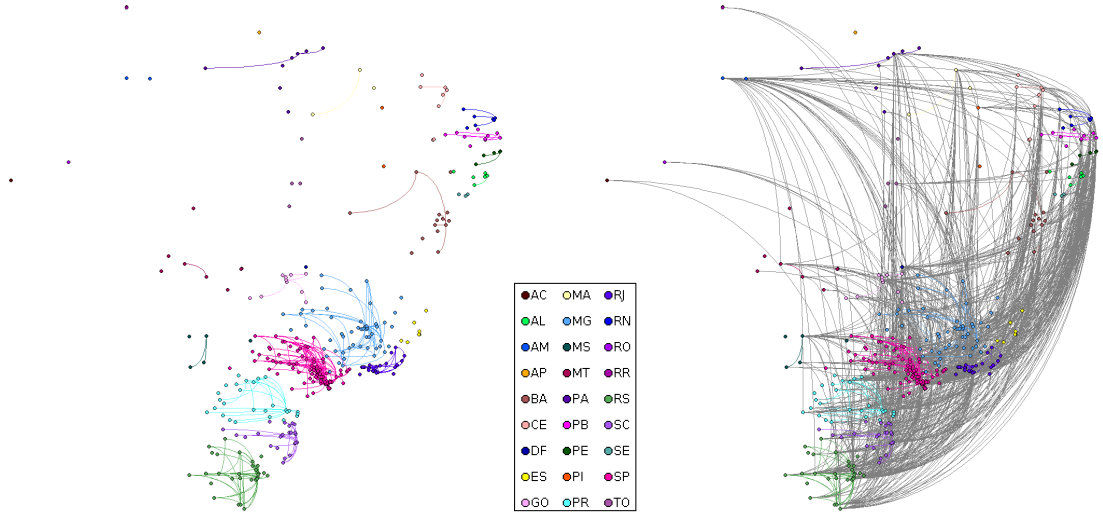

Figura 2. Rede Social dos Doutores em Computação - Cidades

da cidade de sua atuação profissional. Na Figura 4 esses grafos foram reorganizados de forma a destacar as redes estaduais e suas ligações. Para isto foi utilizado um algoritmo que, iterativamente, utiliza uma força de repulsão entre todos os nós (tentando afastar cada nó dos demais) e uma força de atração entre cada par de nós que possua uma aresta (tentando aproximar os nós relacionados). Na imagem da esquerda é possível observar as redes estaduais. Já no grafo da direita é possível observar uma grande concentração de nós ligados em um componente conexo (no centro da imagem).

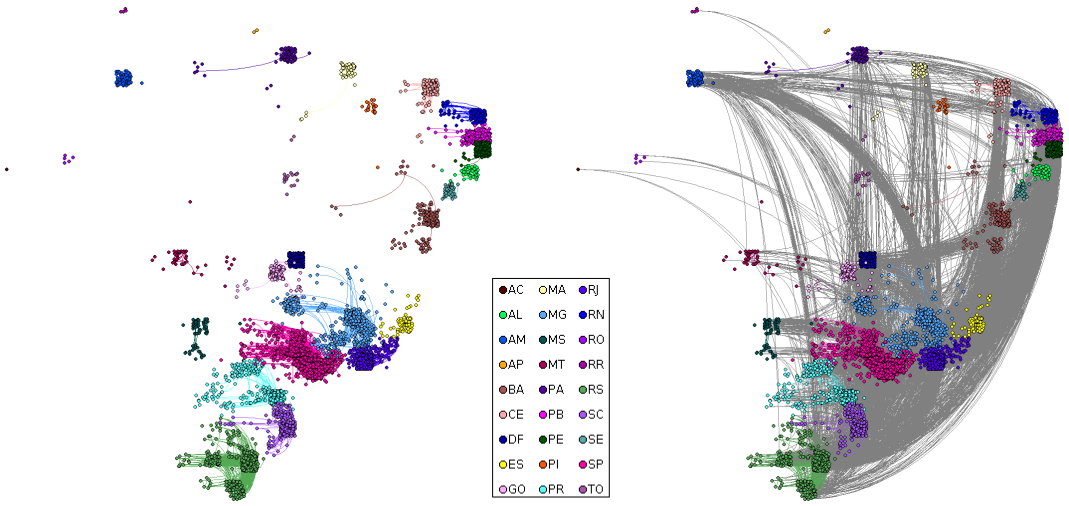

Figura 3. Rede Social dos Doutores em Computação

A Tabela 2 apresenta as métricas da rede nacional e das 27 redes (estaduais e do Distrito Federal). A primeira métrica apresenta o número total de doutores selecionados. Na segunda coluna é possível observar a porcentagem dos nós de cada rede em relação aos 6.358 doutores totais. A maioria dos doutores atuando em computação no país estão em São Paulo, Rio de Janeiro e Minas Gerais (totalizando 53,5% do total). Ao se adicionar os doutores do Rio Grande do Sul, Paraná, Santa Catarina e do Distrito Federal temos mais de 77% do total de doutores. Por outro lado, ao se somar a porcentagem de doutores

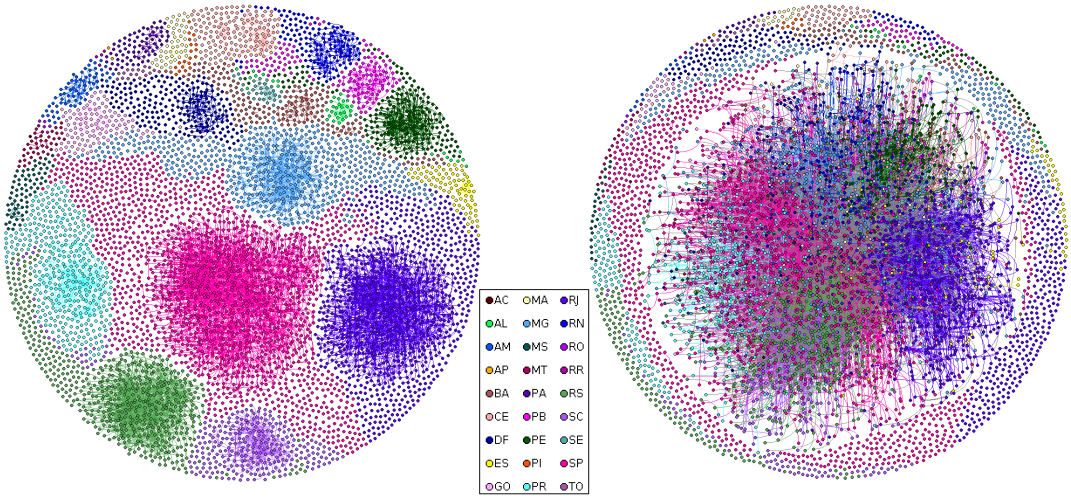

Figura 4. Rede Social dos Doutores em Computação - Reorganizado

do Acre, Amapá, Roraima, Rondônia, Tocantins e Piauí, tem-se menos de 1% do total. Observa-se que as maiores concentrações de doutores ocorre nas cidades (e estados) que possuem programas de pós-graduação na área de ciência da computação[3].

Dos 6.358 doutores estudados, apenas 4.427 possuem alguma ligação com outros doutores da rede (menos de 70% do total). As demais métricas da Tabela 2 consideram apenas estes doutores (chamados na tabela de nós com ligações). Como é possível observar, nenhum dos doutores do Acre, Amapá, Roraima, Rondônia e Tocantins possuem arestas com outros doutores de seus estados, por esta razão as demais métricas destes estados estão com valores em branco nesta tabela.

A coluna *Porcentagem de Nós no Componente Gigante* contém a porcentagem dos nós com ligações que pertencem ao componente gigante (isto é, ao maior componente conexo de cada grafo). É possível observar que 96,9% destes nós estão na componente gigante da rede nacional (4.290 dos 4.427 nós). Os três estados que se destacam nesta métrica são Rio de Janeiro, Pernambuco e Rio Grande do Sul (todos com mais de 90% dos nós com ligações em seus componentes gigantes). Já em Mato Grosso do Sul, Piauí e Bahia este número fica abaixo de 50%. A métrica *densidade* indica quantas arestas existem em um grafo em relação ao número possível de arestas. Nenhum dos grafos correspondentes às redes analisadas possuem alta densidade (a densidade da rede nacional é de aproximadamente 0,001%). As redes mais densas são às do Sergipe, Alagoas e Amazonas. Já as menos densas são de São Paulo, Rio de Janeiro e Minas Gerais.

Na média, cada doutor da rede nacional se relaciona com outros 6,6 doutores. Considerando-se as sub-redes dos estados, os maiores graus médios encontram-se no Rio Grande do Sul, no Rio de Janeiro e no Sergipe. A métrica coeficiente de clusterização costuma ser associada a estabilidade/maturidade de uma rede [Lemieux and Ouimet 2008] pois mede a quantidade de cliques de tamanho 3 (isto é três nós ligados uns aos outros) em relação a quantidade de trios de nós onde haja ao menos duas arestas. Em redes estáveis/maduras espera-se que se o nó A é ligado ao nó B e o nó B é ligado ao nó C então o nó A deverá ser ligado ao nó C (formando um clique contendo estes três nós).

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

Tabela 2. Métricas Calculadas para cara Rede Social Produzida

	Nós Totais	Porcentagem dos Nós	Arestas	Nós com ligações	Nós no Componente Gigante	Porcentagem de Nós no Componente Gigante	Densidade	Grau Médio	Coeficiente de Clusterização	Assortatividade de Grau	Centralização de Grau	Centralização de Closeness	Diâmetro	Tamanho da Clique Máxima	Média dos Caminhos Mínimos
Rede Total	6358	100.00%	14647	4427	4290	96.9%	0.001	6.617	0.048	0.143	0.028	0.000	12	12	4.5
AC	1	0.02%	0	1	-	-	-	-	-	-	-	-	-	-	-
AL	44	0.69%	41	33	26	78.8%	0.078	2.485	0.103	-0.466	0.474	0.564	5	4	2.4
AM	68	1.07%	80	53	43	81.1%	0.058	3.019	0.086	-0.385	0.396	0.138	5	4	2.7
AP	2	0.03%	0	0	-	-	-	-	-	-	-	-	-	-	-
BA	185	2.91%	98	120	51	42.5%	0.014	1.633	0.069	-0.104	0.146	0.015	8	4	3.3
CE	159	2.50%	163	115	98	85.2%	0.025	2.835	0.095	-0.117	0.169	0.097	15	5	5.9
DF	235	3.70%	161	135	87	64.4%	0.018	2.385	0.057	-0.277	0.354	0.049	7	4	3.0
ES	80	1.26%	44	51	26	51.0%	0.035	1.725	0.151	-0.051	0.229	0.050	8	5	3.3
GO	111	1.75%	73	78	50	64.1%	0.024	1.872	0.048	-0.367	0.138	0.073	9	3	3.9
MA	46	0.72%	22	30	15	50.0%	0.051	1.467	0.214	0.339	0.147	0.068	8	4	3.6
MG	610	9.59%	778	432	326	75.5%	0.008	3.602	0.072	-0.071	0.126	0.006	12	7	4.2
MS	68	1.07%	24	44	14	31.8%	0.025	1.091	0.077	-0.051	0.201	0.033	4	3	2.3
MT	42	0.66%	14	25	14	56.0%	0.047	1.120	0.040	-0.571	0.308	0.390	7	3	3.0
PA	73	1.15%	45	52	35	67.3%	0.034	1.731	0.038	-0.344	0.349	0.137	8	3	3.4
PB	140	2.20%	177	106	82	77.4%	0.032	3.340	0.075	-0.159	0.248	0.034	8	4	3.2
PE	222	3.49%	465	177	162	91.5%	0.030	5.254	0.067	-0.279	0.190	0.048	6	5	3.1
PI	20	0.31%	3	11	4	36.4%	0.055	0.545	0.000	-1.000	0.500	1.000	2	2	1.5
PR	397	6.24%	420	289	189	65.4%	0.010	2.907	0.076	-0.011	0.134	0.008	10	5	4.1
RJ	1012	15.92%	1874	704	650	92.3%	0.008	5.324	0.075	-0.065	0.056	0.011	10	6	4.1
RN	141	2.22%	202	115	94	81.7%	0.031	3.513	0.080	-0.257	0.133	0.054	8	4	3.3
RO	5	0.08%	0	2	-	-	-	-	-	-	-	-	-	-	-
RR	4	0.06%	0	1	-	-	-	-	-	-	-	-	-	-	-
RS	564	8.87%	1431	452	409	90.5%	0.014	6.332	0.064	-0.148	0.117	0.011	9	7	3.4
SC	299	4.70%	407	224	183	81.7%	0.016	3.634	0.064	-0.148	0.195	0.019	8	5	3.5
SE	33	0.52%	37	26	22	84.6%	0.114	2.846	0.112	-0.464	0.602	0.682	5	4	2.3
SP	1779	27.98%	2752	1143	984	86.1%	0.004	4.815	0.053	-0.027	0.120	0.002	10	7	4.2
TO	18	0.28%	0	8	-	-	-	-	-	-	-	-	-	-	-

O valor do coeficiente de clusterização pode variar de 0 a 1 e na rede nacional seu valor é de 0,048 (valor baixo tipicamente associado a redes novas/imaturas). O maior valor para esta métrica foi encontrado na rede do Maranhão (0,214) indicando que apesar do número de doutores atuando na área ser relativamente pequeno, há um subgrupo com intensa interação. O segundo maior valor para esta métrica é da rede do Espírito Santo (0,151), seguido pela rede de Sergipe (0,112).

A métrica Assortatividade de Grau verifica a tendência de nós de mesmo grau se relacionarem (o valor -1 indica que não há nós de mesmo grau que se relacionam e o valor 1 indica que todas as arestas ocorrem entre nós de mesmo grau). Ao se analisar a rede nacional, verifica-se que há uma leve tendência de nós de mesmo grau se relacionarem. Já ao se analisar as sub-redes, observa-se que apenas na rede do Maranhão esta tendência ocorre. Um outro tipo de assortatividade que pode ser calculada na rede nacional é assortatividade de estado, isto é, a tendência de pessoas do mesmo estado se relacionarem (o valor -1 indicaria que não existem relações entre pessoas do mesmo estado e o valor 1 indicaria que todas as relações da rede ocorrem entre pessoas do mesmo estado). Na rede nacional o valor desta métrica é de 0.27 indicando uma tendência dos doutores analisados

em ter relações com outros doutores de seus estados.

As medidas de centralização de grau e de *closeness* são baseadas em medidas de centralidade e servem para indicar o quão importante o nó mais central de cada rede é para a sua rede (com base, respectivamente, em seu grau ou nos caminhos médios mínimos entre os nós). Para ambas as métricas, as redes que possuem maior centralização são do Piauí, de Sergipe e de Alagoas indicando que estes três estados são aqueles que possuem indivíduos que mais se destacam em relação aos demais indivíduos de seus estados. Já as redes do Rio de Janeiro, Rio Grande do Sul e de São Paulo são as que apresentam menor centralização de grau e as redes de São Paulo, de Minas Gerais e do Paraná são as que possuem maior centralização de *closeness*.

A métrica diâmetro indica o maior caminho mínimo existente em uma rede (esta métrica está sendo calculada apenas ao componente gigante de cada rede). O diâmetro da rede nacional é de 12, isto é, existe um caminho mínimo necessário para se chegar de um dado nó até outro que passa por 12 pessoas. Redes com diâmetro baixo indicam proximidade entre seus integrantes, as redes com menor diâmetro são de Piauí (diâmetro 2) e de Mato Grosso do Sul (diâmetro 4). Já a rede do Ceará é a que possui maior diâmetro (15). Observa-se que a rede nacional possui menor diâmetro do que a rede do Ceará indicando que ligações entre o Ceará e outros estados encurtam o caminho mínimo entre algumas das pessoas do próprio estado do Ceará.

Uma clique indica a quantidade de nós onde todos os nós estão ligados um ao outro. As redes de Minas Gerais, Rio Grande do Sul e São Paulo possuem cliques máximas de tamanho 7. Já a clique máxima da rede do Piauí possui tamanho 2. Vale destacar que esta rede é bastante pequena. A média dos caminhos mínimos (métrica calculada apenas nos componentes gigantes) indica, na média, qual é o tamanho do caminho necessário para ligar dois nós arbitrários do componente gigante. Na rede nacional, a distância média entre duas pessoas é de 4,5 pessoas.

A seguir, é realizada uma breve análise das arestas existentes na rede nacional. As Tabelas 3 e 4 apresentam, respectivamente, o total de arestas entre cada par de estados e a quantidade relativa destas arestas em relação ao total de arestas de cada estado.

Observa-se nas últimas linhas da Tabela 3 a porcentagem de arestas que são intra-estado (entre duas pessoas do mesmo estado) e inter-estados. Os doutores do Acre, de Roraima, de Rondônia e de Tocantins só possuem arestas inter-estados. Os estados que se destacam por arestas intra-estado são Rio Grande do Sul, São Paulo e Rio de Janeiro, todos os outros estados possuem mais arestas inter-estados do que intra-estados.

Cada linha da Tabela 4 indica a porcentagem de arestas do estado correspondente a esta linha em relação a cada estado (totalizando 100% por linha). De um modo geral, os maiores valores são encontrados na diagonal principal (indicando a tendência de duas pessoas do mesmo estado se relacionarem).

Focando nas porcentagens de fora da diagonal principal, destaca-se: 23% das arestas que envolvem pessoas do Amazonas envolvem pessoas de Minas Gerais; 22% das arestas da Bahia são com São Paulo; 23% das relações do Ceará são com o Rio de Janeiro; 26% das relações do Espírito Santo são com o Rio de Janeiro; 30% das relações de Goiânia ocorrem com São Paulo; 55% das relações do Mato Grosso do Sul são com São Paulo; bem como 40% das relações do Mato Grosso; a maior porcentagem de relações

Tabela 3. Quantidade de Arestas de Acordo com o Estado de Atuação

	AC	AL	AM	AP	BA	CE	DF	ES	GO	MA	MG	MS	MT	PA	PB	PE	PI	PR	RJ	RN	RO	RR	RS	SC	SE	SP	TO
AC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0
AL	0	41	5	0	1	4	10	0	1	0	28	0	0	1	32	28	0	1	38	7	0	0	4	5	1	31	1
AM	0	5	80	0	2	7	1	4	4	0	74	1	1	8	10	21	1	5	37	1	0	0	26	2	7	30	0
AP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BA	0	1	2	0	98	3	7	10	1	1	20	1	0	1	10	46	1	7	39	23	0	0	17	13	5	89	1
CE	0	4	7	0	3	163	3	1	2	1	18	1	0	5	12	33	5	9	102	9	0	0	17	4	0	45	0
DF	0	10	1	0	7	3	161	0	8	1	17	4	0	1	29	21	2	12	55	15	0	0	24	24	4	82	2
ES	0	0	4	0	10	1	0	44	0	1	10	0	0	2	5	2	0	2	41	0	0	0	9	7	1	20	0
GO	0	1	4	0	1	2	8	0	73	1	21	6	1	0	6	5	0	3	48	6	0	0	4	2	0	82	0
MA	0	0	0	0	1	1	1	1	1	22	5	0	1	0	9	1	3	1	23	0	0	0	6	1	1	28	0
MG	0	28	74	0	20	18	17	10	21	5	778	2	3	5	28	30	1	61	251	25	1	2	76	40	3	352	2
MS	0	0	1	0	1	1	4	0	6	0	2	24	0	0	0	0	0	10	27	0	0	0	6	3	0	103	0
MT	0	0	1	0	0	0	0	0	1	1	3	0	14	0	0	0	0	0	6	0	0	0	1	3	1	21	0
PA	0	1	8	0	1	5	1	2	0	0	5	0	0	45	3	8	0	3	10	7	0	0	17	9	1	20	0
PB	0	32	10	0	10	12	29	5	6	9	28	0	0	3	177	91	1	9	56	48	0	0	21	5	2	54	1
PE	0	28	21	0	46	33	21	2	5	1	30	0	0	8	91	465	7	13	57	66	0	0	35	6	27	120	3
PI	0	0	1	0	1	5	2	0	0	3	1	0	0	0	1	7	3	0	3	12	0	0	6	1	0	0	0
PR	0	1	5	0	7	9	12	2	3	1	61	10	0	3	9	13	0	420	51	18	0	0	68	122	3	301	3
RJ	0	38	37	0	39	102	55	41	48	23	251	27	6	10	56	57	3	51	1874	92	0	0	166	34	8	334	3
RN	0	7	1	0	23	9	15	0	6	0	25	0	0	7	48	66	12	18	92	202	0	0	54	7	8	77	0
RO	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0
RR	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
RS	2	4	26	0	17	17	24	9	4	6	76	6	1	17	21	35	6	68	166	54	0	0	1431	195	6	228	1
SC	0	5	2	0	13	4	24	7	2	1	40	3	3	9	5	6	1	122	34	7	0	0	195	407	1	79	2
SE	0	1	7	0	5	0	4	1	0	1	3	0	1	1	2	27	0	3	8	8	0	0	6	1	37	15	0
SP	1	31	30	0	89	45	82	20	82	28	352	103	21	20	54	120	0	301	334	77	4	1	228	79	15	2752	0
TO	0	1	0	0	1	0	2	0	0	0	2	0	0	0	1	3	0	3	3	0	0	0	1	2	0	0	0
Intra	0%	17%	24%	0%	25%	37%	33%	28%	27%	21%	42%	13%	27%	31%	29%	43%	7%	37%	56%	30%	0%	0%	59%	42%	28%	57%	0%
Inter	100%	83%	76%	0%	75%	63%	67%	72%	73%	79%	58%	87%	73%	69%	71%	57%	93%	63%	44%	70%	100%	100%	41%	58%	72%	43%	100%

existentes entre a Paraíba é com o Ceará, porém o Ceará está mais proporcionalmente mais relacionado com São Paulo do que com a Paraíba; 26% das relações do Piauí ocorrem com o Rio de Janeiro; 80% das relações de Rondônia são com São Paulo; 67% das relações de Roraima ocorrem com Minas Gerais; 20% das relações de Santa Catarina ocorrem com o Rio Grade de Sul; 21% das relações de Sergipe ocorrem com Pernambuco; e as pessoas de São Paulo estão, proporcionalmente, mais relacionadas com pessoas de Minas Gerais e do Rio de Janeiro.

A última análise realizada utilizou mineração de textos para, a partir dos títulos dos artigos publicados pelos doutores de cada região, identificar as principais palavras-chave utilizadas. Ao todo foram identificados 92.580 artigos diferentes (em periódicos e em anais de eventos) publicados em inglês. Optou-se por utilizar apenas artigos publicados em inglês pois a análise realizada é baseada na frequência de expressões. Foram consideradas expressões de uma, duas e três palavras. Devido à limitação de espaço, serão apresentados apenas os resultados referentes a expressões de duas palavras[4].

O processamento dos títulos envolveu a remoção de *stop-words*, eliminação dos títulos repetidos dentro de um mesmo estado, e contagem da frequência de todas as expressões de uma, duas ou três palavras (para as 28 redes). Por fim foi verificada a frequência relativa de cada expressão em cada estado relação a frequência relativa da mesma expressão na rede nacional, de forma a identificar quais expressões são relativamente mais utilizadas em cada estado. Nos resultados serão apresentadas apenas expressões cuja frequência na rede nacional seja de pelo menos 50 ocorrências.

[4]Expressões de duas palavras costumam ser mais significativas para o entendimento dos assuntos aos quais as publicações se referem.

Tabela 4. Porcentagem de Arestas de Acordo com o Estado de Atuação

	AC	AL	AM	AP	BA	CE	DF	ES	GO	MA	MG	MS	MT	PA	PB	PE	PI	PR	RJ	RN	RO	RR	RS	SC	SE	SP	TO
AC	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	67%	0%	0%	33%	0%
AL	0%	17%	2%	0%	0%	2%	4%	0%	0%	0%	12%	0%	0%	0%	13%	12%	0%	0%	16%	3%	0%	0%	2%	2%	0%	13%	0%
AM	0%	2%	24%	0%	1%	2%	0%	1%	1%	0%	23%	0%	0%	2%	3%	6%	0%	2%	11%	0%	0%	0%	8%	1%	2%	9%	0%
AP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BA	0%	0%	1%	0%	25%	1%	2%	3%	0%	0%	5%	0%	0%	0%	3%	12%	0%	2%	10%	6%	0%	0%	4%	3%	1%	22%	0%
CE	0%	1%	2%	0%	1%	37%	1%	0%	0%	0%	4%	0%	0%	1%	3%	7%	1%	2%	23%	2%	0%	0%	4%	1%	0%	10%	0%
DF	0%	2%	0%	0%	1%	1%	33%	0%	2%	0%	4%	1%	0%	0%	6%	4%	0%	2%	11%	3%	0%	0%	5%	5%	1%	17%	0%
ES	0%	0%	3%	0%	6%	1%	0%	28%	0%	1%	6%	0%	0%	1%	3%	1%	0%	1%	26%	0%	0%	0%	6%	4%	1%	13%	0%
GO	0%	0%	1%	0%	0%	1%	3%	0%	27%	0%	8%	2%	0%	0%	2%	2%	0%	1%	18%	2%	0%	0%	1%	1%	0%	30%	0%
MA	0%	0%	0%	0%	1%	1%	1%	1%	1%	21%	5%	0%	1%	0%	8%	1%	3%	1%	22%	0%	0%	0%	6%	1%	1%	26%	0%
MG	0%	2%	4%	0%	1%	1%	1%	1%	1%	0%	42%	0%	0%	0%	2%	2%	0%	3%	14%	1%	0%	0%	4%	2%	0%	19%	0%
MS	0%	0%	1%	0%	1%	1%	2%	0%	3%	0%	1%	13%	0%	0%	0%	0%	0%	5%	14%	0%	0%	0%	3%	2%	0%	55%	0%
MT	0%	0%	2%	0%	0%	0%	0%	0%	2%	2%	6%	0%	27%	0%	0%	0%	0%	0%	12%	0%	0%	0%	2%	6%	2%	40%	0%
PA	0%	1%	5%	0%	1%	3%	1%	1%	0%	0%	3%	0%	0%	31%	2%	5%	0%	2%	7%	5%	0%	0%	12%	6%	1%	14%	0%
PB	0%	5%	2%	0%	2%	2%	5%	1%	1%	1%	5%	0%	0%	0%	29%	15%	0%	1%	9%	8%	0%	0%	3%	1%	0%	9%	0%
PE	0%	3%	2%	0%	4%	3%	2%	0%	0%	0%	3%	0%	0%	1%	8%	43%	1%	1%	5%	6%	0%	0%	3%	1%	2%	11%	0%
PI	0%	0%	2%	0%	2%	11%	4%	0%	0%	7%	2%	0%	0%	0%	2%	15%	7%	0%	7%	26%	0%	0%	13%	2%	0%	0%	0%
PR	0%	0%	0%	0%	1%	1%	1%	0%	0%	0%	5%	1%	0%	0%	1%	1%	0%	37%	5%	2%	0%	0%	6%	11%	0%	27%	0%
RJ	0%	1%	1%	0%	1%	3%	2%	1%	1%	1%	7%	1%	0%	0%	2%	2%	0%	2%	56%	3%	0%	0%	5%	1%	0%	10%	0%
RN	0%	1%	0%	0%	3%	1%	2%	0%	1%	0%	4%	0%	0%	1%	7%	10%	2%	3%	14%	30%	0%	0%	8%	1%	1%	11%	0%
RO	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	20%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	80%	0%
RR	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	67%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	33%	0%
RS	0%	0%	1%	0%	1%	1%	1%	0%	0%	0%	3%	0%	0%	1%	1%	1%	0%	3%	7%	2%	0%	0%	59%	8%	0%	9%	0%
SC	0%	1%	0%	0%	1%	0%	2%	1%	0%	0%	4%	0%	0%	1%	1%	1%	0%	13%	3%	1%	0%	0%	20%	42%	0%	8%	0%
SE	0%	1%	5%	0%	4%	0%	3%	1%	0%	1%	2%	0%	1%	1%	2%	21%	0%	2%	6%	6%	0%	0%	5%	1%	28%	11%	0%
SP	0%	1%	1%	0%	2%	1%	2%	0%	2%	1%	7%	2%	0%	0%	1%	2%	0%	6%	7%	2%	0%	0%	5%	2%	0%	57%	0%
TO	0%	5%	0%	0%	5%	0%	11%	0%	0%	0%	11%	0%	0%	0%	5%	16%	0%	16%	16%	0%	0%	0%	5%	11%	0%	0%	0%

A Tabela 5 apresenta as cinco expressões de duas palavras mais frequentes em todas as redes apresentadas. É possível notar que as três expressões mais usadas na rede nacional são *neural networks*, *real time* e *case study*. Esta tabela ajuda a ilustrar quais assuntos são mais frequentes nos títulos das publicações para cada estado.

Já a Tabela 6 apresenta as cinco expressões de duas palavras cujas frequências relativas são maiores em cada estado quando comparadas com a frequência relativa destas expressões na rede nacional. Esta tabela ajuda a identificar as peculiaridades dos assuntos publicados em cada um dos estados. É possível observar que em Alagoas há destaque para alguns tópicos de inteligência artificial; em Goiás há assuntos relacionados à bioinformática; no Mato Grosso do Sul destacam-se assuntos relacionados à computação paralela/distribuída; e no Paraná há destaque para alguns tópicos de inteligência artificial.

4. Trabalhos Correlatos

Nos últimos anos, diversos trabalhos analisaram redes sociais acadêmicas. Os trabalhos mais relacionados ao presente trabalho são aqueles que utilizaram dados da Plataforma Lattes ou da DBLP[5] para analisar e comparar redes acadêmicas.

A área de ciência da computação foi analisada [Menezes et al. 2009] através da rede de colaborações comparando-se a produtividade e as características de algumas regiões do mundo (Brasil, América do Norte e Europa).

Os dados bibliográficos da DBLP foram utilizados para o estudo e análise da rede de colaborações em ciência da computação [Franceschet 2011]. Os resultados deste artigo indicam que as colaborações que originam publicações em periódicos são mais fortes do que aqueles que resultam em publicações em anais de eventos.

[5]http://dblp.uni-trier.de/db/

Tabela 5. Expressões Mais Frequentes em Cada Estado

	Expressão 1	Expressão 2	Expressão 3	Expressão 4	Expressão 5
Rede Total	neural networks	real time	case study	neural network	finite element
AC	-	-	-	-	-
AL	wireless sensor	sensor networks	multi agent	intelligent tutoring	learning environment
AM	sensor networks	wireless sensor	real time	hard real	embedded systems
AP	-	-	-	-	-
BA	pierre auger	auger observatory	product lines	software product	real time
CE	wireless sensor	sensor networks	dielectric properties	glass ceramics	decision analysis
DF	hoc networks	quantum dots	neural networks	case study	quantum wells
ES	context aware	neural networks	conceptual modeling	ontology based	neural network
GO	multi objective	case study	context aware	evolutionary algorithm	neural networks
MA	agent based	multi agent	neural networks	power systems	lung nodule
MG	sensor networks	wireless sensor	neural networks	real time	schistosoma mansoni
MS	mato grosso	grosso sul	coarse grained	covered graphs	matching covered
MT	fuzzy logic	neural networks	power quality	case study	citizenship community
PA	performance evaluation	case study	wireless networks	genetic algorithms	corynebacterium pseudotuberculosis
PB	virtual reality	real time	petri nets	object oriented	petri net
PE	neural networks	time series	real time	neural network	petri net
PI	optical networks	reconfigurable architecture	general purpose	h264 avc	noc based
PR	genetic algorithm	neural networks	real time	wireless sensor	genetic algorithms
RJ	finite element	neural networks	multi agent	real time	case study
RN	real time	aspect oriented	brain machine	multi agent	neural networks
RO	essential oil	hpaec pad	-	-	-
RR	-	-	-	-	-
RS	real time	multi agent	case study	object oriented	h264 avc
SC	real time	neural networks	multi agent	neural network	wireless sensor
SE	aedes aegypti	larvicidal activity	particle swarm	swarm optimization	essential oil
SP	neural networks	neural network	real time	case study	artificial neural
TO	dimensional electron	electron gas	magnetic field	antidot lattice	classical magnetoresistance

Informações da DBLP também foram utilizadas em um estudo sobre a rede social acadêmica brasileira em ciência da computação [Freire and Figueiredo 2011]. Este trabalho identificou que existem alguns *super peers* na rede (indivíduos que se destacam por possuírem um grau muito grande), bem como compararam algumas métricas da rede com a nota da avaliação CAPES atribuída aos programas de pós-graduação.

Com um objetivo parecido, foi realizado um trabalho combinando informações dos currículos Lattes, com citações do Google Scholar e do Microsoft Academic Search para a realização de uma análise profunda da produtividade dos programas de pós-graduação em ciência da computação brasileiros comparando-se diversas métricas de bibliométricas e de redes sociais a fim de classificar os programas usando diferentes perspectivas. Por fim, neste trabalho foram correlacionadas as métricas analisadas com a nota da avaliação feita pela CAPES e foi possível identificar um claro padrão na classificação.

A principal característica que diferencia o trabalho atual dos trabalhos correlatos é a análise da rede social de ciência da computação no Brasil como um todo (e não apenas dos docentes pertencentes aos programas de pós-graduação), incluindo uma análise das redes estaduais, combinada com a identificação das expressões mais frequentes na produção de cada um dos estados a fim de se identificar peculiaridades.

5. Conclusões

Este artigo apresentou um análise geral das redes sociais de doutores que atuam na área de ciência da computação no Brasil. Estas redes foram formadas exclusivamente com

Tabela 6. Expressões Relativamente Mais Frequentes em Cada Estado

	Expressão 1	Expressão 2	Expressão 3	Expressão 4	Expressão 5
AC	-	-	-	-	-
AL	intelligent tutoring	network design	tutoring systems	coloured petri	semantic web
AM	hard real	wireless mesh	based testing	scheduling problems	interactive digital
AP	-	-	-	-	-
BA	pierre auger	auger observatory	product lines	software product	mobile agent
CE	alzheimer disease	user interfaces	optical properties	parallel programming	sensitivity analysis
DF	quantum dots	quantum wells	mobile hoc	hoc networks	face recognition
ES	fault diagnosis	context aware	ontology based	agent oriented	business process
GO	structure prediction	evolutionary algorithm	problem solving	protein structure	quality service
MA	intrusion detection	power flow	semi automatic	parallel distributed	power systems
MG	digital library	minas gerais	schistosoma mansoni	digital libraries	network design
MS	parallel algorithm	parallel applications	peer peer	power flow	parallel processing
MT	electric power	fuzzy logic	hardware implementation	mobile hoc	intrusion detection
PA	multi user	wireless networks	wireless mesh	mesh networks	brazilian amazon
PB	coloured petri	virtual reality	petri nets	interval valued	petri net
PE	series forecasting	hard real	optical networks	petri net	energy consumption
PI	noc based	optical networks	h264 avc	model based	performance evaluation
PR	genetic programming	paraconsistent logic	manufacturing systems	knowledge discovery	policy based
RJ	square root	nuclear power	element methods	rio janeiro	power plant
RN	multi user	interval valued	spanning tree	aspect oriented	minimum spanning
RO	-	-	-	-	-
RR	-	-	-	-	-
RS	white dwarf	graph grammars	h264 avc	video coding	motion estimation
SC	suspension polymerization	intrusion detection	mobile agent	expert systems	network management
SE	communication systems	hardware software	natural language	swarm optimization	particle swarm
SP	optimum path	path forest	langmuir blodgett	blodgett films	collisions root
TO	project management	tree problem	quantum wells	collaborative learning	-

informações extraídas dos currículos Lattes destes doutores.

A análise realizada, envolvendo tanto a rede nacional dos doutores como também as sub-redes estaduais considerou tanto características dos grafos de relacionamento quanto a produção bibliográfica dos docentes. Constatou-se que as redes são bastante heterogêneas; pouco densas; e que provavelmente ainda estão novas/em processo de formação. Estas características podem ser justificadas pelo fato da área de ciência da computação ser uma área nova; com assuntos bastante diversificados e também pela diversidade (cultural e social) e amplitude do país.

Por outro lado os componentes gigantes de cada uma das redes estaduais e da rede nacional são compostos pela grade maioria dos nós que possuem ao menos uma conexão, mostrando uma boa conectividade da rede, cuja média dos caminhos mínimos é de 4,5 indicando que, na média, todos os doutores pertencentes ao componente gigante estão relativamente próximos na rede.

Como trabalhos futuros pretende-se aprofundar a análise da rede de doutores em ciência da computação, bem como realizar uma análise semelhante envolvendo todos os doutores que atuam no Brasil.

Agradecimentos

O trabalho apresentado neste artigo foi parcialmente financiado pelo Programa de Educação Tutorial do MEC, FAPESP, CNPq, CAPES e Pró-reitoria de graduação da USP.

Referências

Brandão, M. A., Moro, M. M., Lopes, G. R., and Oliveira, J. P. (2013). Using link semantics to recommend collaborations in academic social networks. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 833–840, Republic and Canton of Geneva, Switzerland.

Digiampietri, L., Mena-Chalco, J., de Jésus Pérez-Alcázar, J., Tuesta, E. F., Delgado, K., and Mugnaini, R. (2012a). Minerando e caracterizando dados de currículos Lattes. In *CSBC-BraSNAM 2012*.

Digiampietri, L., Mena-Chalco, J., Silva, G. S., Oliveira, L., Malheiro, A., and Meira, D. (2012b). Dinâmica das relações de coautoria nos programas de pós-graduação em computação no Brasil. In *CSBC-BraSNAM 2012*.

Digiampietri, L. A., Mena-chalco, J. P., Melo, P. O. V., Malheiros, A. P., Meira, D. N. O., Franco, L. F., and Oliveira, L. B. (2014). BraX-Ray: An X-Ray of the Brazilian Computer Science Graduate Programs. *Plos-One*, (no prelo):20.

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.

Freire, V. and Figueiredo, D. (2011). Ranking in collaboration networks using a group based metric. *Journal of the Brazilian Computer Society*, pages 1–12.

Gao, S., Denoyer, L., and Gallinari, P. (2012). Link prediction via latent factor blockmodel. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 507–508, New York, NY, USA. ACM.

Hew, K. F. (2011). Review: Students' and teachers' use of facebook. *Comput. Hum. Behav.*, 27(2):662–676.

Hsieh, C.-J., Tiwari, M., Agarwal, D., Huang, X. L., and Shah, S. (2013). Organizational overlap on social networks and its applications. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 571–582, Republic and Canton of Geneva, Switzerland.

Lemieux, V. and Ouimet, M. (2008). *Análise Estrutural das Redes Sociais*. Instituto Piaget.

Melo-Minardi, R., Digiampietri, L., de Melo, P. O. V., Jr., G. F., and Oliveira, L. (2013). Caracterização dos programas de pós-graduação em bioinformática no Brasil. In *CSBC-BraSNAM 2013*.

Menezes, G. V., Ziviani, N., Laender, A. H. F., and Almeida, V. (2009). A geographical analysis of knowledge production. In *in Computer Science In Proceedings of the 18th international conference on World Wide Web*, pages 1041–1050.

Miyata, B., Kano, V., and Digiampietri, L. (2013). Combinando mineração de textos e análise de redes sociais para a identificação das áreas de atuação de pesquisadores. In *CSBC-BraSNAM 2013*.

Xu, Y., Guo, X., Hao, J., Ma, J., Lau, R. Y. K., and Xu, W. (2012). Combining social network and semantic concept analysis for personalized academic researcher recommendation. *Decis. Support Syst.*, 54(1):564–573.