Programas de mestrado em Matemática no Brasil: uma aplicação de redes para caracterizar seus títulos

Master's degree programs in Mathematics in Brazil: an application of networks to characterize their titles

Programas de maestría en matemáticas en Brasil: una aplicación de redes para caracterizar sus títulos

8

ARTÍCULO

Inácio de Sousa Fadigas

Departamento de Ciências Exatas
Universidade Estadual de Feira de Santana

Possui graduação em Engenharia Civil pela Universidade Estadual de Feira de Santana (1984) e mestrado em Engenharia Civil (Geotecnia) pela Universidade Federal da Paraíba (1987) e Doutorado em Difusão do Conhecimento (UFBA/UEFS/LNCC/UNEB/IFBA/CIMATEC 2011). É professor Titular da Universidade Estadual de Feira de Santana. Tem também especialização em Educação Matemática (UEFS 1998) e experiência na área de Matemática. Coordenou o Núcleo de Educação Matemática Omar Catunda da UEFS de 1997 a 2008. Atualmente participa do grupo Fuxicos e Boatos, dedicado a estudos e pesquisas em Ciência de Redes.

fadigas@uefs.br
orcid.org/0000-0002-9330-506X

Trazíbulo Henrique Pardo Casas

Departamento de Ciências Exatas
Universidade Estadual de Feira de Santana
Possui graduação em Engenharia Civil pela Universidade Estadual de Feira de Santana (1984), mestrado em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro (1991) e doutorado em Informática na Educação pela Universidade Federal do Rio Grande do Sul (2003). Atualmente é professor adjunto b da Universidade Estadual de Feira de Santana. Tem experiência na área de Educação, com ênfase em Informática na Educação, atuando principalmente nos seguintes temas: informática, informática na educação, educação matemática e educação, ciência cognitiva. cibercultura, filosofia e tecnologia, difusão do conhecimento.

henrique@uefs.br
orcid.org/0000-0003-1756-530X

Marcos Grilo Rosa

Departamento de Ciências Exatas
Universidade Estadual de Feira de Santana

Possui graduação em Licenciatura Em Matemática pela Universidade Estadual de Feira de Santana (2001), mestrado em Matemática pela Universidade Federal de Pernambuco (2004) e doutorado em Difusão do Conhecimento pela Universidade Federal da Bahia (2016). Atualmente é professor adjunto da Universidade Estadual de Feira de Santana. Tem experiência em Geometria e Topologia, Teoria de Redes, Teoria dos Grafos e Ensino de Matemática.

grilo@uefs.br
orcid.org/0000-0002-6382-3907

Hernane Borges de Barros Pereira

Departamento de Educação
Universidade do Estado da Bahia e SENAI CIMATEC
Hernane Borges de Barros Pereira concluiu o doutorado em Engenharia Multimídia pela Universitat Politècnica de Catalunya em 2002. Atualmente é Professor Pleno do Departamento de Educação da Universidade do Estado da Bahia e Professor Associado do SENAI CIMATEC. É docente do Programa de Pós-graduação em Modelagem Computacional e Tecnologia Industrial e do Doutorado Multi-institucional e Multidisciplinar em Difusão do Conhecimento. É consultor ad-hoc do Ministério da Educação. Dentro do âmbito da ciência e tecnologia da informação e inovação, trabalha com temas como: redes sociais e complexas, difusão do conhecimento, engenharia de software, interação homem-computador, etc. usando técnicas de modelagem computacional.

hbbpereira@gmail.com
orcid.org/0000-0001-7476-9267

RECEBIDO: 10 de janeiro de 2020 / ACEITO: 26 de janeiro de 2020

Resumo:

O trabalho aborda redes dos títulos das dissertações de mestrados em matemática, no Brasil. A análise usa redes semânticas de títulos (RST) para caracterizar qualitativa e quantitativamente as redes. Foram selecionados 41 cursos de mestrado e construídas as RST usando-se a abordagem de redes por cliques, na qual as palavras dos títulos são mutuamente conectadas. Foi aplicado o método Louvain (algoritmo) para detectar as comunidades de palavras. As redes foram caracterizadas e agrupadas por região geográfica, o que levou a inferir distinções entre regiões com base em regularidade ou não dos grupos de palavras.

PALAVRAS-CHAVE

Redes semânticas, Comunidades, Títulos, Dissertações, Disciplinaridade.

Abstract

This paper dealt with the networks of titles of master's dissertations in mathematics in Brazil. Semantic Networks of Titles (SNT) are analyzed to characterize networks qualitatively and quantitatively. 41 master's courses were selected and the

SNTs were built using a network-based approach by cliques, in which the words of the titles are mutually connected. Louvain's method (algorithm) was applied to detect communities of words. The networks were characterized and grouped by geographical region, which led to infer distinctions between regions depending on whether the word groups were regular or not.

KEYWORDS

Semantic networks, Dissertations, Communities, Titles, Dissertations, Disciplines.

Resumen:
El trabajo realiza un análisis de redes de los títulos de diversos proyectos de maestrías en matemáticas de Brasil. El análisis utiliza redes de títulos semánticos (RST) caracterizados cualitativa y cuantitativamente. Para ello se seleccionaron 41 cursos de capacitación y utilizando el enfoque de redes de clic, en el que las palabras de los enlaces están conectadas. El método de Louvain (algoritmo) se aplicó para detectar comunidades de palabras. Las redes fueron caracterizadas y agrupadas por región y área geográfica, lo que condujo a inferencias de regiones basadas en regularidad o no, de los grupos de palabras.

PALABRAS CLAVE:

Redes semánticas, Comunidades, Títulos, Disertaciones, Disciplina.

1. INTRODUÇÃO

Os programas ou cursos de mestrado em Matemática no Brasil se fazem presentes em todas as unidades da Federação. A cada ano são produzidas dezenas de dissertações, cujos títulos dizem respeito às diversas áreas de pesquisa da Matemática. As palavras que formam os títulos, quando analisadas em conjunto, trazem em si informações que permitem caracterizar a(s) temática(s) principais daquele programa/curso. A análise feita aqui usa redes semânticas de títulos (RST) para caracterizar quantitativamente as redes e suas possíveis interpretações fornecem inserções qualitativas. A técnica está fundamentada na abordagem de redes por cliques, devida principalmente a Pereira et al. (2011).

Não encontramos na literatura pesquisas que tratam em particular das redes de títulos das dissertações dos mestrados em Matemática no Brasil, o que a torna particularmente original. Para realizar a caracterização, foram usados índices gerais da teoria das redes sociais e complexas e um algoritmo para detecção de comunidades.

Por tratar-se de um país de dimensões continentais, também foram investigadas as semelhanças e diferenças das redes de títulos das dissertações, agrupando os programas/cursos de acordo com a região geográfica do Brasil na qual a instituição sede está localizada. Esta investigação é suscitada pelo fato de tais regiões apresentarem diferenças geopolíticas, populacionais, econômicas, culturais e educacionais significativas, o que poderia influenciar também nas linhas/áreas de investigação científica da Matemática.

O texto foi estruturado em 5 tópicos principais: *Referencial Teórico*, no qual é fundamentada a base da pesquisa, e inclui os subtópicos sobre Redes Semânticas de Títulos, Detecção de Comunidades em Redes; *Procedimentos Metodológicos*, que inclui a Coleta de Dados, Construção das Redes, Aplicação do Método Louvain; *Resultados e Discussões*, no qual constam Redes

por Curso e Redes dos Cursos Agrupados por Região Geográfica e *Considerações Finais*.

2. REFERENCIAL TEÓRICO

2.1. REDES SEMÂNTICAS DE TÍTULOS

A teoria das redes complexas, ou a ciência das redes, tem uma vasta aplicação, em particular no campo semântico, quando trata de redes nas quais os vértices são palavras e as arestas são ligações entre os vértices, determinadas por alguma propriedade. Nos trabalhos de Caldeira et al. (2006) e Teixeira et al. (2010), que construíram e analisaram redes de palavras em discursos, os vértices são as palavras que formam as sentenças e as arestas ligam mutuamente os vértices. Nos trabalhos de Fadigas et al. (2009), Pereira et al. (2011), Cunha et al. (2013), Henrique et al. (2014) e Grilo et al. (2017), os quais estudaram redes de títulos de artigos publicados em periódicos, as arestas ligam mutuamente os vértices que pertencem ao mesmo título. Na presente pesquisa, segue-se o mesmo padrão: os vértices da rede são as palavras dos títulos das dissertações de mestrado em matemática do Brasil, mutuamente conectados pelas arestas. Assim, dois títulos distintos são conectados se apresentarem uma ou mais palavras em comum.

2.2. DETECÇÃO DE COMUNIDADES EM REDES

O termo comunidade, na perspectiva de redes, remete intuitivamente a subgrupos cujos laços (arestas) conectando internamente o grupo é mais denso que as arestas que conectam os grupos distintos (Murata, 2010). O conceito de comunidades como grupos coesivos tem sua origem na análise estrutural, que busca identificar a conectividade de indivíduos dentro e entre os grupos (Wellman, 1997).

A detecção de comunidades em redes não é uma tarefa simples, principalmente quando se trata de grandes redes. Apenas para citar alguns trabalhos, Girvan e Newman (2002) propõem um método para encontrar comunidades usando a ideia de índices de centralidade para delimitar o contorno da comunidade. Para redes complexas, Capocci et al. (2005) desenvolveram um algoritmo para detectar comunidades baseado nos métodos espectrais e leva em conta o peso nas arestas e a orientação das ligações. Nesta mesma linha, Clauset, Newman e Moore (2004) apresentam um algoritmo de aglomeração hierárquica que pretende ser rápido na detecção de comunidades em grandes redes.

O objetivo principal da detecção de comunidades é separar uma rede em grupos de vértices com poucas conexões entre os mesmos. Para tal objetivo, uma medida usada é a modularidade, que identifica o contraste na densidade de arestas dentro do grupo, comparada com o valor esperado para uma distribuição aleatória de arestas, ou seja, mede a qualidade de cada partição.

Uma expressão para computar um número que expresse a modularidade em uma rede pode ser encontrada, por exemplo, em Barabási (2016), e é dada por

$$M = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right] \qquad (1)$$

Na expressão, n_c é o número de comunidades, L_c é o número de arestas na comunidade, L é o número total de arestas e k_c é o número total de graus dos vértices na comunidade. Quanto

mais alto o valor de *M* (que não excede 1), melhor é a partição da estrutura de comunidade. O valor de *M* = 0 expressa que a rede total é uma comunidade simples. Modularidade negativa ocorre se cada vértice pertence a comunidades separadas. Para exibir comunidades de palavras nas redes de títulos das dissertações, estudadas no presente trabalho, optamos pelo Método Louvain (Blondel et al., 2008), que dispõe de um algoritmo rápido e com boa precisão, cuja detecção das comunidades é baseada na medida da modularidade. O algoritmo implementado no Pajek (Batagelj e Mrvar, 1998) introduz um parâmetro de resolução *r*, que torna a Equação 1 em:

$$M = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - r \left(\frac{k_c}{2L} \right)^2 \right] \quad (2)$$

3. PROCEDIMENTOS METODOLÓGICOS

3.1 COLETA DE DADOS

A etapa preliminar da coleta de dados consistiu em relacionar os cursos recomendados pela Capes (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) vinculada ao Ministério da Educação (MEC) do Brasil, após consulta ao sítio oficial http://www.capes.gov.br/ (acessado em setembro de 2013), limitando a busca à Grande Área de "Ciências Exatas e da Terra" na Área de Avaliação "Matemática (Matemática/Probabilidade e Estatística)". Dos 45 cursos encontrados no sítio foram selecionados os 42 que são específicos de Matemática. Na etapa seguinte, verificamos que em um dos cursos ainda não havia dissertações defendidas, e o número final ficou em 41, conforme discriminado na Tabela 1. Para a construção das redes, a fonte principal para a obtenção dos dados foram os Cadernos Indicadores da CAPES, no sítio http://conteudoweb.capes.gov.br/conteudoweb/CadernoAvaliacaoServlet. No sítio encontram-se os Cadernos do período de 1998 até 2012, no qual as avaliações dos programas eram feitas por triênio, que serviu de base para a coleta. Com a mudança da avaliação por quadriênio, não consideramos a coleta de 2016. Escolhido o ano de interesse e a instituição desejada, buscou-se a área a ser pesquisada, no caso, Matemática/Probabilidade e Estatística. O grupo intitulado "TE-Teses e Dissertações" quando acionado exibe um documento no formato PDF com vários campos delimitados, dos quais foram coletados [autor]; [título]; [linha de pesquisa]. Os dados de todos os campos foram transferidos para uma planilha, organizados com um campo para cada coluna. O procedimento foi feito para cada curso, de 1998 até 2012.

3.2 CONSTRUÇÃO DAS REDES

As redes de palavras dos títulos das dissertações foram construídas a partir da planilha de dados. Cada título é copiado para uma linha em um arquivo de texto, que é submetido a dois tipos de tratamento: o primeiro consiste em retirar do texto sinais de pontuação como vírgulas, pontos, hífens e outros sinais. O segundo é mais complexo e se utiliza uma rotina desenvolvida por Caldeira (2005) a qual faz uso de três programas: (i) o pacote UNITEX, disponível em http://www-igm.univ-mlv.fr/ unitex/; (ii) o programa Ambisin, desenvolvido por Caldeira (2005) para tratar questões como a eliminação de ambiguidades, eliminação de palavras gramaticais e separação das formas flexionadas ou canônicas das palavras; (iii) o programa NetPal, desenvolvido pelo prof. Doutor José Garcia Vivas Miranda e seus colaboradores, que gera a rede no formato apropriado para o uso do programa Pajek, criado por Vladimir Batagelj e

n.	Nome do curso	IES	UF
1	Matemática	FUFPI	PI
2	Matemática	IMPA	RJ
3	Matemática	PUC-RIO	RJ
4	Matemática Aplicada e Computacional	UEL	PR
5	Matemática	UEM	PR
6	Matemática	UFABC	SP
7	Matemática	UFAL	AL
8	Matemática	UFAM	AM
9	Matemática	UFBA	BA
10	Matemática	UFC	CE
11	Matemática	UFCG	PB
12	Matemática	UFES	ES
13	Matemática	UFF	RJ
14	Matemática	UFG	GO
15	Matemática	UFJF	MG
16	Matemática	UFMA	MA
17	Matemática	UFMG	MG
18	Matemática e Estatística	UFPA	PA
19	Matemática	UFPB_JP	PB
20	Matemática	UFPE	PE
21	Matemática Aplicada	UFPR	PR
22	Matemática	UFRGS (ma)	RS
23	Matemática Aplicada	UFRGS (maplic)	RS
24	Matemática	UFRJ (ma)	RJ
25	Matemática Aplicada	UFRJ (maplic)	RJ
26	Matemática Pura e Aplicada	UFSC	SC
27	Matemática	UFSCar	SP
28	Matemática	UFSM	RS
29	Matemática	UFU	MG
30	Matemática	UFV	MG
31	Matemática	UNB	DF
32	Matemática Aplicada e Computacional	UNESP_PP	SP
33	Matemática Universitária	UNESP_RC	SP
34	Matemática	UNESP_SJRP (ma)	SP
35	Matemática Aplicada	UNESP_SJRP (maplic)	SP
36	Matemática	UNICAMP (ma)	SP
37	Matemática Aplicada e Computacional	UNICAMP (macomp)	SP
38	Matemática Aplicada	UNICAMP (maplic)	SP
39	Matemática	USP (ma)	SP
40	Matemática Aplicada	USP (maplic)	SP
41	Matemática	USP/SC	SP

Tabela 1. Relação dos cursos utilizados na pesquisa

Tabela com o nome do curso, Instituição de Ensino Superior (IES) ao qual está vinculado e Unidade Federativa (UI) da unidade da Instituição.

Andrej Mrvar (Batagelj e Mrvar, 1998). O tabela 1 mostra os cursos que foram pesquisados.

Após a construção das redes para cada curso, os arquivos foram agrupados por região geográfica na qual está instalada a sede de cada programa ou curso. Este recorte permite investigar características regionais na escolha dos títulos das dissertações.

3.3. APLICAÇÃO DO MÉTODO LOUVAIN

O algoritmo consiste em otimizar a modularidade, de forma a encontrar a partição que resulta em um maior valor da modularidade. A descrição completa do algoritmo, aqui resumida, é encontrada em (Blontel *et al*. 2008). O algoritmo é baseado em dois passos repetidos iterativamente. No primeiro passo é considerado que todos os vértices formam sua própria

comunidade, ou seja, temos *N* vértices formando *N* comunidades. No segundo passo, é feita uma busca ordenada em todos os vértices de 1 até *N*, de forma que o vértice vizinho é incorporado àquela comunidade se existe um crescimento na modularidade. Este passo é realizado iterativamente até que a máxima modularidade local seja atingida. Nesse passo cada vértice pode ser visitado várias vezes. Uma vez que o máximo local tenha sido atingido, o algoritmo constrói uma nova rede na qual os vértices são as comunidades encontradas, com os pesos dos laços entre as comunidades calculado como a soma dos pesos totais entre os vértices das comunidades. O segundo passo é repetido iterativamente (sobre a rede cujos vértices representam comunidades), o que conduz a uma decomposição hierárquica da rede. Isto torna o algoritmo apropriado para tratar grandes redes.

3.3.1. IMPLEMENTAÇÃO DO ALGORITMO NO PAJEK

Na página de Andrej Mrvar na Internet, http://mrvar.fdv.uni-lj.si/pajek/community/Community-DrawExample.htm, é encontrada uma descrição de como utilizar o método Louvain implementado no Pajek. Sem entrar em detalhes técnicos inerentes ao programa, destacamos apenas algumas informações que julgamos pertinentes:

1. São disponibilizadas duas formas de acessar rotinas distintas do algoritmo. Uma, denominada Multi-Level Coarsening + Single Refinement, realiza somente o refinamento da partição obtida no último nível (uma partição mais grosseira). A outra, chamada de Multi-Level Coarsening + Multi-Level Refinement, difere da primeira por realizar fases grosseiras e refinadas para cada nível obtido.

2. É recomendado tentar o algoritmo para diferentes valores do parâmetro de resolução *r* (Equação 2), cujo padrão é 1. Altos valores de resolução produzem grandes números de comunidades, enquanto baixos valores (maiores que 0) produzem poucas comunidades.

3. Para obter resultados melhores, mesmo sem a maximização da modularidade, sugere-se comparar as partições obtidas em duas rodadas de execução do algoritmo com o mesmo parâmetro de resolução para avaliar a correlação. O próprio programa Pajek dispõe de rotinas para esta avaliação: o V de Cramer, o Rajski e o Índice ajustado de Rand. Se a correlação entre as duas partições é pequena, o número de comunidades provavelmente não é o correto, e por isso deve-se tentar o algoritmo com um outro valor (maior ou menor) do parâmetro de resolução *r*. É sugerido também o uso do índice de correlação mais alto encontrado, mesmo que a modularidade não seja a maior.

Estes procedimentos foram usados para as redes de palavras dos títulos das dissertações, separadas por regiões geográficas, bem como para a rede total, ou seja, para a rede com todos os títulos que formam o componente maior conectado.

4. RESULTADOS E DISCUSSÕES

4.1 REDES POR CURSO

A Tabela 2 mostra as principais quantidades encontradas para caracterizar as redes de palavras de títulos de cada curso.

Como se pode inferir da Tabela 2, cerca de metade dos cursos (21 cursos) produziram dissertações em todo período de coleta. Dos 20 restantes, dois deles, de Matemática Aplicada, foram encerrados antes de 2012 (da UFRJ e da UNESP-SJRP), e 18 cursos começaram após 1998. Estas quantidades traduzem o cresci-

IES	Período	*QT*	*QV*	*QA(p=1)*	*QA(p>1)*	*QC*	*% MC*	*DR*
FUFPI	2010-2012	21	93	309	39	5	51,61	0,370
IMPA	1998-2012	51	235	918	72	8	87,23	0,408
PUC-RIO	1998-2012	108	380	1472	103	3	98,16	0,349
UEL	2009-2012	16	95	394	37	3	86,32	0,565
UEM	2001-2012	101	255	984	260	7	86,32	0,285
UFABC	2009-2012	20	94	289	8	4	72,34	0,746
UFAL	2005-2012	50	173	564	50	7	87,28	0,405
UFAM	2002-2012	57	208	752	65	6	88,94	0,385
UFBA	1998-2012	136	353	1299	142	3	98,02	0,330
UFC	1998-2012	163	369	1513	213	3	98,10	0,285
UFCG	2004-2012	75	248	1115	175	4	93,55	0,307
UFES	2008-2012	28	124	405	40	7	77,42	0,533
UFF	1998-2012	93	275	784	82	7	94,18	0,367
UFG	1998-2012	146	402	1699	244	4	97,26	0,293
UFJF	2011-2012	8	37	112	6	3	51,35	0,585
UFMA	2012-2012	4	18	63	1	3	61,11	1,000*
UFMG	1998-2012	145	400	1492	179	5	95,75	0,198
UFPA	2005-2012	109	396	2021	359	1	100,00	0,294
UFPB_JP	1998-2012	179	452	1934	341	1	100,00	0,279
UFPE	1998-2012	108	334	1143	80	5	96,11	0,351
UFPR	2004-2012	42	164	541	46	5	92,68	0,425
UFRGS (ma)	1998-2012	127	405	1635	166	5	96,30	0,267
UFRGS (maplic)	1998-2012	205	618	3483	516	2	98,87	0,198
UFRJ (ma)	1998-2012	121	352	1259	186	6	96,31	0,341
UFRJ (maplic)	1998-2009	48	184	621	41	4	86,41	0,415
UFSC	1998-2012	94	303	1253	150	6	95,38	0,330
UFSCar	1998-2012	102	315	1218	172	6	94,29	0,323
UFSM	2008-2012	24	112	435	36	1	100,00	0,558
UFU	2009-2012	28	133	485	38	5	84,96	0,375
UFV	2009-2012	23	83	205	20	5	84,34	0,488
UNB	1998-2012	198	566	2649	374	4	98,76	0,240
UNESP_PP	2012-2012	7	52	246	4	1	100,00	0,553
UNESP_RC	2010-2012	38	108	303	20	5	91,67	0,442
UNESP_SJRP (ma)	1998-2012	167	418	1637	220	5	96,89	0,250
UNESP_SJRP (maplic)	1999-2006	70	218	972	128	1	100,00	0,309

UNICAMP (ma)	1998-2012	188	463	1743	187	6	97,84	0,244
UNICAMP (maplic)	1998-2012	137	519	2487	195	3	98,46	0,262
UNICAMP (ma-comp)	2007-2012	65	251	1215	94	6	91,63	0,266
USP (ma)	1998-2012	150	411	1380	131	3	98,54	0,291
USP (maplic)	1998-2012	87	330	1345	66	3	97,58	0,370
USP/SC	1998-2012	142	358	1433	199	5	96,93	0,296

Tabela 2. Quantitativos para as redes semânticas de títulos por curso

Nota: *QT*- quantidade de títulos; *QV*- quantidade de vértices; *QA(p*=1) - quantidade de arestas com peso 1; *QA(p*>1) - quantidade de arestas com peso maior que 1; %*MC* - percentagem do maior componente (em termos de número de vértices); *DR*- diâmetro de referência.

(*) A rede com diâmetro de referência igual a 1 é um caso singular da RST UFMA que tem apenas 4 títulos, 3 dos quais formam o maior componente como uma clique.

mento da oferta de mestrados em Matemática no período.

A quantidade de títulos varia em uma faixa que vai de 4 a 205 títulos, ou seja, há uma variação grande nestas quantidades, influenciada sobretudo pelo período de funcionamento do curso. Assim, por exemplo, o curso de mestrado em Matemática da UFMA apresenta apenas 4 títulos porque foi implantado em 2012, data final da coleta. Por outro lado, a variação na quantidade de títulos para cursos em funcionamento desde o início (1998) até o fim (2012) da coleta, distingue a produção dos mesmos.

O tamanho de cada rede, dado pela quantidade de vértices, mostra que, do ponto de vista das redes complexas, não são redes grandes, sendo a maior com 618 vértices (UFRGS – Matemática Aplicada). O tamanho médio dos títulos, que é a relação entre o número de vértices *(n)* e o número de títulos (n_q), varia entre 2,26 e 7,43. Essa relação revela a maior ou menor diversidade de palavras na escolha dos títulos. Comparando-se, por exemplo, a RST-IMPA que possui 51 títulos e 253 palavras distintas, com a RST-UFC, que possui 163 títulos e 369 palavras distintas, a análise da relação n/n_q (RST-IMPA: 4,61; RST-UFC: 2,26) mostra que um programa pode ter menos títulos que outro mas, proporcionalmente, o seu vocabulário é mais diversificado quando comparado com a quantidade de títulos utilizada.

A Tabela 2 também traz informações sobre a quantidade de ligações entre os vértices, isto é, a quantidade de ligações estabelecidas entre as palavras dos títulos para cada curso. Como é considerada a quantidade de vezes em que ocorre a ligação entre duas palavras, as colunas 3 e 4 diferenciam ligações que ocorreram apenas uma vez (peso 1) das ligações que ocorreram mais de uma vez (peso maior que 1). Os resultados mostram a predominância de ocorrência de ligações com peso 1, ou seja, a maioria das ligações entre as palavras ocorre apenas uma vez.

Um dos parâmetros que indica o quanto a rede está fragmentada, isto é, quantos grupos de palavras não têm ligações entre si, é a quantidade de componentes. Nota-se da Tabela 2 que das 41 redes, apenas 5 delas (cerca de 12%) têm apenas 1 componente. Este dado seria um indicativo da maior ou menor diversidade na escolha de grupos de palavras para compor os títulos, porém os percentuais dos maiores componentes, que indicam a quantidade de

palavras destes em relação à rede total, mostram que os menores componentes não são significativos para esta interpretação. De fato, em 29 cursos (70,7%), observa-se que o tamanho do maior componente está acima de 90% de vértices. Observa-se também que das redes restantes, ou seja, com o maior componente com menos que 90% de vértices, todas são incompletas em relação ao período coletado (1998 a 2012). Há, pois, uma tendência de crescimento deste percentual. A relação entre a quantidade de títulos e o tamanho percentual do maior componente das RST é apresentada na Figura 1.

A Figura 1, mostra que há uma tendência de que redes com maior número de títulos exibam maiores componentes com tamanhos próximos a 100%. Assim, se duas redes com quantidades de títulos equiparáveis apresentam diferença significativa no tamanho do maior componente, isto pode ser atribuído à escolha de temáticas distintas na RST com menor percentual do componente maior. Por exemplo, a RST UEM (101 títulos; maior componente 86,32%), comparada com a RST UFScar (102 títulos; maior componente 94,29%) pode ser dita apresentar temáticas mais diversificadas. Esta indicação é associada qualitativamente ao fato de que o programa da UEM tem destaque em três áreas (álgebra, análise e geometria) enquanto o programa da UFScar apresenta destaque em apenas duas áreas (análise e geometria).

Outro parâmetro de rede que pode ser usado para caracterizar os programas ou cursos é o diâmetro da rede. O diâmetro de uma rede conectada é o comprimento do maior caminho geodésico entre um par de vértices. No caso das RST conectadas, representa o afastamento máximo entre duas palavras presentes em títulos distintos. Portanto, diz respeito à diversidade de temas usados na formação dos títulos.

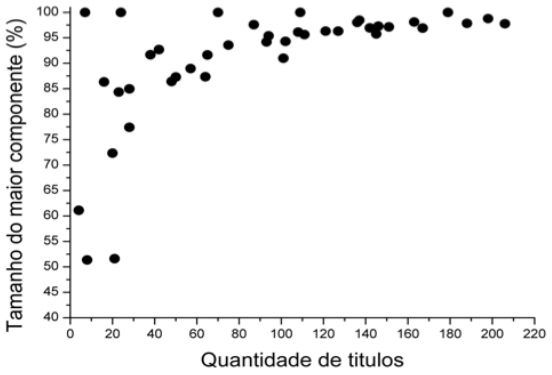

Figura 1. Gráfico da distribuição do tamanho do maior componente em relação a quantidade de títulos.

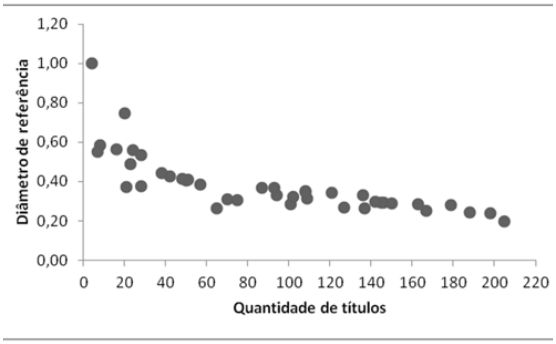

Figura 2. Gráfico da variação do diâmetro de referência com a quantidade de títulos para as RST.

Conforme definem Fadigas e Pereira (2013), o diâmetro de referência é uma normalização do diâmetro relativo ao maior diâmetro possível em uma estrutura de cliques minimamente conectada, ou seja, cada clique é conectada a outra por um único vértice. Esta estrutura é denominada "estrutura em linha". Para as RST, a Tabela 2 mostra que apenas 1% destas está na menor faixa do diâmetro de referência (0 a 0,25), que corresponde a uma "estrutura em estrela" e, portanto, a uma rede mais coesiva, no sentido de aproximação geodésica entre seus títulos. Contudo, o diâmetro de referência é influenciado pelo tamanho da rede, ou seja, pela quantidade de títulos. Para possibilitar comparações, a Figura 2 mostra a variação do diâmetro de referência com o tamanho dos

títulos para as RST. De acordo com a Figura 2, há uma tendência de redução do diâmetro de referência com o aumento da quantidade de títulos. Assim, os valores maiores do diâmetro de referência são mais influenciados pela menor quantidade de títulos do que por uma possível diversidade de vocabulário, por exemplo (Figura 2).

4.2. REDES DOS CURSOS AGRUPADOS POR REGIÃO GEOGRÁFICA

4.2.1. QUANTITATIVOS GERAIS

Para termos uma abordagem macro, tanto quantitativa quanto qualitativa, das redes de títulos das dissertações em Matemática, agrupamos as redes por região geográfica do Brasil, o que permitiu analisar as diferenças e semelhanças entre as redes, e inferir alguma característica regional. A exemplo, as informações da Tabela 2 permitem inferir que os programas situados na Região Sudeste foram os que mais produziram dissertações (1826), seguida da Região Nordeste (736), da Região Sul (609), da Região Centro-Oeste (344) e da Região Norte (166). É interessante notar porém que a média de dissertações por região, quando calculada pela quantidade de cursos, apresenta a Região Centro-Oeste como a que mais produz (172 por curso), enquanto que as outras regiões apresentam uma média entre 83 e 92 dissertações por curso, sendo que a Região Norte aparece como a menos produtiva (83 por curso), empatada com a Região Sudeste. Quando o cálculo leva em conta os períodos de cada curso, a Região Sudeste apresenta um pequeno aumento em relação à Região Norte (7,97 contra 7,90), enquanto as demais regiões continuam na mesma ordem de produção. O fato de a Região Centro-Oeste se apresentar como a mais produtiva está ligado fortemente à produção da UnB (Tabela 3).

O número de vértices da rede indica a quantidade de palavras distintas usadas no conjunto dos títulos, ou seja, excluída dessa quantidade as palavras repetidas. Para quantificar as repetições de palavras quando as redes de cada programa individualmente são juntadas para formar as redes por região, introduzimos um índice que quantifica esta redução de vértices em comum. Este índice é calculado pela Equação 3, na qual S_{no} é a soma dos vértices das redes de cada programa e n é o número total de vértices da rede de programas daquela região.

$$IRV = \frac{S_{n_0} - n}{S_{n_0}} \qquad (3)$$

A Tabela 3 apresenta algumas quantidades para as redes por região, na qual observa-se que a rede da Região Sudeste é a que apresenta o *IRV* mais alto. O valor (55,2%) indica que mais da metade das palavras usadas nos títulos forma um vocabulário de fonte comum a todos os cursos daquela região. Quando se leva em conta a quantidade de títulos, a parametrização indicada pelo *IRVp* mostra que a RST da Região Sudeste apresenta o menor índice, enquanto as outras apresentam índices na mesma ordem de grandeza. Isso significa que há uma maior "disciplinaridade" em termos de vocabulário escolhido para os títulos na Região Sudeste.

4.2.2. COMPONENTES E COMUNIDADES DE PALAVRAS

A informação da quantidade de componentes de cada rede por região, aliado ao percentual

Rede da região ...	*NCurso*	*QT*	*NComp*	*%MC*	*n*	*IRV*	*IRVp*
Centro-Oeste	2	344	4	98,991	793	0,180	0,053
Nordeste	8	736	3	99,674	1228	0,398	0,054
Norte	2	166	2	99,457	552	0,086	0,052
Sudeste	22	1826	14	98,492	1295	0,552	0,030
Sul	7	609	5	98,996	2569	0,337	0,055

Tabela 3: Caracterização das redes com destaque para os componentes

Nota: *NCurso* - número de cursos; *QT*- quantidade de títulos; *NComp*- número de componentes; *%MC* - percentagem do maior componente (em termos de número de vértices); *n*- número de vértices da rede; *IRV*- índice de redução de vértices; *IRVp* - índice de redução de vértices parametrizado (*IRV/QT*).

de vértices da maior componente (Tabela 3) já indica o comportamento em relação a grupos de palavras que são mais próximas daquela separadas de outros grupos de palavras. Assim, a presença de mais de um componente na rede significa que há grupos de palavras isoladas, por alguma razão que é peculiar a cada rede. Conforme mostra a Tabela 3, o tamanho da maior componente para todas as redes está em torno de 99%, o que mostra a predominância da maior componente sobre as menores, ou seja, para todas as redes há uma predominância de um grande grupo de palavras ligadas entre si. Portanto, não há diferença considerável na topologia das redes quanto ao número de componentes e a distribuição de tamanhos das componentes, o que indica uma uniformidade na escolha do vocabulário usado nos títulos das dissertações, em cada grupo de curso por região.

Apesar da análise da quantidade e da distribuição dos tamanhos dos componentes fornecer uma primeira aproximação em termos de grupos, esta não apresenta detalhes sobre as características do maior componente. O objetivo da aplicação do algoritmo para detecção de comunidades nas redes de palavras dos títulos das dissertações por região geográfica é detectar similaridades e dissimilaridades entre o maior componente das redes e inferir interpretações para os resultados. As comunidades de palavras são caracterizadas por apresentarem um número de ligações internas maior do que aquele esperado para uma distribuição aleatória de arestas na rede.

Para uma análise mais precisa da formação de comunidades de palavras, foram construídas redes apenas com os vértices da maior componente conectada, que neste caso é representativo das redes como um todo, tendo em vista os percentuais das maiores componentes, conforme a Tabela 3. O algoritmo do Método Louvain (Blondel et al., 2008) está incorporado ao programa Pajek (Batagelj e Mrvar, 1998), aqui usado na pesquisa das comunidades. A sua aplicação para se obter resultados mais confiáveis requer a investigação da resolução *r* ótima, parâmetro que afeta diretamente o valor da modularidade e em consequência, o número de comunidades: menores valores resultam em uma quantidade menor de comunidades.

A modularidade, por sua vez, indica a qualidade na divisão das comunidades, de forma que altos valores (máximo de 1) implicam em ligações densas entre os vértices dentro do grupo e ligações esparsas entre vértices de grupos dis-

Rede da região ...	*R*	*M_r*	Índice V Cramer	Comunidades	*%MComun*
Centro-Oeste	0,25	0,751963	1,000000	3	96,56
Nordeste	0,25	0,750069	1,000000	2	99,67
Norte	0,25	0,815102	0,974503	3	45,72
Sudeste	0,10	0,900062	1,000000	2	98,89
Sul	1,50	0,399884	0,759705	19	9,91

Tabela 4: Quantitativos relativos à detecção de comunidades

tintos. Porém, a modularidade isoladamente não resulta em uma determinação precisa da quantidade de grupos na rede, pois uma mesma resolução pode resultar em modularidades e quantidades de grupos distintas, quando da execução do algoritmo. Seguindo as orientações descritas na subseção 2.3, referentes à implementação no Pajek e aplicação nas redes, foram feitos os testes no sentido de otimizar a aplicação do algoritmo. Para cada rede analisada, foram usados os valores de resoluções 0,1; 0,2; 0,25; 0,5; 1,0; 1,25; 1,5 e 1,75 e escolhida a resolução que resultou em um maior valor do índice estatístico conhecido como V de Cramer, que mede a correlação no cálculo das modularidades para pares de resoluções iguais. As colunas de 2 a 4 da Tabela 4 trazem os resultados para as redes por região (Tabela 4).

Observa-se também da Tabela 4 que, do ponto de vista de grupos ou comunidades de palavras, as redes podem ser divididas em três categorias: na primeira categoria, inclui-se as redes do Centro-Oeste, Nordeste e Sudeste, com um número baixo de comunidades (2 ou 3) e uma comunidade com quase todos os vértices (maior que 96,5%). No segundo grupo, está a rede da região Norte, com apenas 3 comunidades, mas com os tamanhos das comunidades bem distribuídos (as outras duas comunidades menores têm percentuais de 16,39% e 37,89%). Finalmente, na terceira categoria, está a rede da Região Sul, com 19 comunidades, de tamanhos bem distribuídos entre 1,79% e 9,91% (maior comunidade inclusa). Vale ressaltar que esta última rede não é muito estável, e tomando-se a resolução 1,5 que resultou em um maior valor do índice V de Cramer, o número de comunidades ainda pode variar. Para as demais redes, foram encontrados os mesmos números de comunidades nas várias execuções do algoritmo para a resolução ótima.

Do ponto de vista das palavras usadas nos títulos, as redes das regiões Nordeste, Sudeste e Centro-Oeste comportam-se como um grupo coeso, no qual as ligações entre as palavras dentro de uma comunidade são maiores do que o esperado, caso a distribuição fosse aleatória. Os valores das modularidades entre 0,750 e 0,900 constatam este comportamento. No caso da rede da região Norte, também pode-se inferir esta característica, mas a coesão interna apontada pela modularidade de 0,815 implica na distinção de três grupos de palavras, devido a certa homogeneidade nos tamanhos dos grupos. A rede da região Sul apresenta-se com o valor mais baixo da modularidade (0,395), indicando uma mais fraca coesão interna dos grupos, ou seja, as ligações são próximas daquelas esperadas para uma distribuição aleatória de ligações entre as palavras. Assim, do ponto de vista da modularidade, a rede de palavras da região Sudeste é a que apresenta uma melhor qualidade de divisão de comunidades, enquanto a rede da região Sul é a que

apresenta a mais fraca divisão. Se considerarmos o número de cursos para estas duas últimas regiões, observa-se um contraste: os 22 cursos da região Sudeste apresentam-se mais homogêneos em termos de escolha das palavras dos títulos, em contraste com os 7 cursos da região Sul que são mais heterogêneos.

5. CONSIDERAÇÕES FINAIS

A análise feita a partir das redes de palavras dos títulos das dissertações de Mestrado para cada um dos programas/cursos revela que há uma diversidade de períodos de funcionamento, quantidade de títulos, quantidade de vértices, quantidade de aresta com peso, percentual do maior componente e diâmetro de referência. Tal diversidade, porém, não impede que sejam feitas inferências sobre o comportamento das redes a longo prazo, a exemplo do crescimento assintótico do percentual do maior componente (tende para 100%) com o aumento da quantidade de título e, portanto, da quantidade de palavras (vértices) que os formam.

O comportamento do diâmetro de referência com o crescimento da quantidade de títulos mostra que há uma tendência em todas as redes de que à medida que novos títulos forem acrescidos, a distância geodésica máxima entre duas palavras na rede diminua.

Quando a comparação é feita com as RST agrupadas por região geográfica, observa-se que há também uma diferenciação no aspecto da produção acadêmica, refletida pela quantidade de títulos das regiões, com destaque para a Região Sudeste, em termos absolutos. Porém, o fato da Região Centro-Oeste contar com apenas duas instituições pesquisadas, dentre elas uma instituição notadamente produtiva como é o caso da UnB, a destaca em média de produtividade. Em termos do vocabulário usado nos títulos, a rede dos títulos das dissertações da Região Sudeste é menos diversificada, o que indica uma maior "disciplinaridade". A aplicação do método Louvain para detecção de comunidades mostrou-se apropriada para diferenciar as redes com maior ou menor diversidade de coesão entre grupos de palavras que formam os títulos. Destaca-se o comportamento da rede da Região Sul que apresenta maior diversidade de coesão, expressa pela maior quantidade de comunidades e menor modularidade.

O uso do algoritmo Louvain para detectar comunidades mostrou-se satisfatório quando empregado em redes semânticas de títulos, pois consegue identificar e quantificar os grupos de palavras com maior coesão interna, cuja interpretação inicialmente está ligada à disciplinarização. Este método pode ser explorado junto com outras análises quantitativas e qualitativas que leve em conta as linhas de pesquisa de cada curso/programa, associadas às áreas definidas por órgãos de fomento a pesquisa científica e tecnológica, a exemplo do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), no Brasil, de forma a identificar temas que são de interesses regionais ou nacional; se estes estão em consonância com o que se estuda no mundo; se há focos de estudos característico no país que justifiquem vocabulários mais ou menos específicos.

BIBLIOGRAFIA

Barabási, A. L. (2016). *Network science*. Cambridge university press.

Batagelj, V. e Mrvar, A. (1998). Pajek-program for large network analysis. *Connections*, 21(2), 47-57.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., e Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008.

Caldeira, S. M. G. (2005). *Caracterização da Rede de Signos Linguísticos: Um modelo baseado no aparelho psíquico de Freud*. Master in Computer Modeling, Fundação Visconde de Cairu.

Caldeira, S. M., Lobao, T. P., Andrade, R. F. S., Neme, A., e Miranda, J. V. (2006). The network of concepts in written texts. *The European Physical Journal B-Condensed Matter and Complex Systems*, 49(4), 523-529. doi:10.1140/epjb/e2006-00091-3

Capocci, A., Servedio, V. D., Caldarelli, G., e Colaiori, F. (2005). Detecting communities in large networks. *Physica A: Statistical Mechanics and its Applications*, 352(2-4), 669-676. doi:10.1016/j.physa.2004.12.050

Clauset, A., Newman, M. E. J., e Moore, A. (2004). Finding community structure in very large networks. Physical review E, 70(6), 066111.

Cunha, M. V., Rosa, M. G., Fadigas, I. S., Miranda, J. G. V., e Pereira, H. B. B. (2013, August). Redes de títulos de artigos científicos variáveis no tempo. In *Anais do II Brazilian Workshop on Social Network Analysis and Mining* (pp. 194-205). SBC.

Fadigas, I. S., Henrique, T., Senna, V., Moret, M. A., e Pereira, H. B. B. (2009). Análise de redes semânticas baseada em títulos de artigos de periódicos científicos: o caso dos periódicos de divulgação em educação matemática.

Fadigas, I. S. e Pereira, H. B. B. (2013). A network approach based on cliques. *Physica A: Statistical Mechanics and its Applications*, 392(10), 2576-2587. doi:10.1016/j.physa.2013.01.055

Girvan, M. e Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences, 99*(12), 7821-7826. doi:10.1073/pnas.122653799

Grilo, M., Fadigas, I. S., Miranda, J. G. V., Cunha, M. V., Monteiro, R. L. S., e Pereira, H. B. B. (2017). Robustness in semantic networks based on cliques. *Physica A: Statistical Mechanics and its Applications*, 472, 94-102. doi:10.1016/j.physa.2016.12.087

Henrique, T., Fadigas, I. S., Rosa, M. G., e Pereira, H. B. B. (2014). Mathematics education semantic networks. *Social Network Analysis and Mining,* 4(1), 200. doi:10.1007/s13278-014-0200-x

Murata, T. (2010). Detecting communities in social networks. In *Handbook of social network technologies and applications* (pp. 269-280). Springer, Boston, MA. doi:10.1007/978-1-4419-7142-5_12

Pereira, H. B. B., Fadigas, I. S., Senna, V., e Moret, M. A. (2011). Semantic networks based on titles of scientific papers. *Physica A: Statistical Mechanics and its Applications,* 390(6), 1192-1197. doi:10.1016/j.physa.2010.12.001

Teixeira, G. M., Aguiar, M. S. F. D., Carvalho, C. F., Dantas, D. R., Cunha, M. V., Morais, J. H. M., ... e Miranda, J. G. V. (2010). Complex semantic networks. *International Journal of Modern Physics C*, 21(03), 333-347. doi:10.1142/S0129183110015142

Wellman, B. (1997). Structural analysis: From method and metaphor to theory and substance. *Contemporary Studies in Sociology*, 15, 19-61.