# 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***

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

8

ARTICLE

**Inácio de Sousa Fadigas**

Department of Exact Sciences
Feira de Santana State University

He holds a degree in Civil Engineering from the State University of Feira de Santana (1984), a master's degree in Civil Engineering (Geotechnics) from the Federal University of Paraíba (1987) and a PhD in Knowledge Dissemination (UFBA / UEFS / LNCC / UNEB / IFBA / CIMATEC 2011). He is a full time professor at the Feira de Santana State University. He also has a specialization in Mathematics Education (UEFS 1998) and experience in Mathematics. He coordinated the Omar Catunda Mathematics Education Center of UEFS from 1997 to 2008. He currently participates in the group Fuxicos e Boatos, dedicated to studies and research in Network Science.

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

**Trazíbulo Henrique**

Department of Exact Sciences
Feira de Santana State University

Graduated in Civil Engineering from the Feira de Santana State University (1984), master's degree in Systems and Computer Engineering from the Federal University of Rio de Janeiro (1991) and has a doctorate in Informatics in Education from the Federal University of Rio Grande do Sul (2003). He is currently an adjunct professor at the Feira de Santana State University. He has experience in education with emphasis on Informatics in Education, acting mainly on the following subjects: informatics, informatics in education, mathematical education, cognitive Science, cyberculture, philosophy, technology and dissemination of knowledge.

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

## Marcos Grilo Rosa

Department of Exact Sciences
Feira de Santana State University

He has a degree in Mathematics from the Feira de Santana State University (2001), a master's degree in Mathematics from the Federal University of Pernambuco (2004) and a doctorate in Dissemination of Knowledge from the Federal University of Bahia (2016). He is currently an adjunct professor at the Feira de Santana State University. He has experience in Geometry and Topology, Network Theory, Graph Theory and Mathematics Teaching.

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

## Hernane Borges de Barros Pereira

Department of Education & Computational Modeling
Bahia State University & Centro Universitário SENAI CIMATEC

He completed his doctorate in Multimedia Engineering at the Universitat Politècnica de Catalunya in 2002. He is currently Full Professor at the Department of Education at the State University of Bahia and Associate Professor at the SENAI CIMATEC. He lectures in the Graduate Program in Computational Modeling and Industrial Technology and, also, in the Graduate Program in Knowledge Diffusion. He serves as an ad-hoc consultant to the Brazilian Ministry of Education. Professor Pereira main interests in the fields of science, information technology and innovation lie in the areas of social and complex networks, diffusion of knowledge, software engineering and human computer interaction, through computer modeling techniques.

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

RECEIVED: January 10, 2020 / ACCEPTED: January 26, 2020

### 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 SNDs were built using a network-based approach by cliques, in which the words of the degrees 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, 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.*

***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.*

## 1. INTRODUCTION

Master's programs or courses in Mathematics in Brazil are present in all units of the Federation. Dozens of dissertations are produced every year, whose titles are related to the various areas of research in Mathematics. The words that make up the titles, when analyzed together, bring with them information that allows characterizing the main themes of that program or course. The analysis performed here uses semantic networks of titles of the master’s dissertations (SNT) to quantitatively characterize the networks and their possible interpretations provide qualitative insertions. The technique is based on the approach of networks by cliques, mainly due to Pereira *et al*. (2011).

In the literature, no research was found that deals particularly with the subject of the networks of titles of the master's dissertations in Mathematics in Brazil, which makes this study particularly original. To carry out the characterization, general indices of the theory of complex and social networks and a community detection algorithm were used.

Being a country of continental dimensions, the similarities and differences of the networks of dissertation titles were also investigated, grouping the programs/courses according to the geographical region of Brazil in which the host institution is located. This research is driven by the fact that such regions have significant geopolitical, population, economic, cultural and educational differences that could also influence the lines or areas of scientific research in Mathematics.

The text is structured in five sections: Section 2 consists of the referential framework in which the theoretical basis of the research is collected and includes the subsections on semantic networks of titles, detection of network communities; Section 3 presents the methodological procedures that include data collection, networking, and the application of the Louvain method; the Section 4 that presents the results and discussions and includes networks by course and course networks grouped by geographic region; finally, in Section 5 presents the concluding remarks.

## 2. THEORETICAL FRAMEWORK

### 2.1. SEMANTIC NETWORKS OF TITLES

The theory of complex networks, or the network science, has a wide application particularly in the semantic field when it deals with to networks in which the vertices are words and the edges are connections between the vertices, determined by some property. In the works of Caldeira *et al*. (2006) and Teixeira *et al*. (2010), who built and analyzed networks of words in written and oral discourses, the vertices are the words that form the sentences and the edges mutually connect the vertices. In the works of Fadigas *et al*. (2009), Pereira *et al*. (2011), Cunha *et al*. (2013), Henrique *et al*. (2014) and Grilo *et al*. (2017), who studied networks of article titles published in journals, the edges mutually connect the vertices that belong to the same title. In this research, the same pattern is followed: the vertices of the network are the words of the titles of the master's dissertations of mathematics in Brazil, connected to each other by the edges. Therefore, two different titles are connected if they have one or more words in common.

### 2.2. NETWORK COMMUNITY DETECTION

The term community, from the network perspective, refers to subgroups in which the edges that internally connect the group make it denser than the edges that connect to different groups (Murata, 2010). The concept of communities, as cohesive groups, has its origin in the structural analysis that seeks to identify the connectivity of individuals within and between groups (Wellman, 1997).

Detecting communities in networks is not a simple task, especially when it comes to large networks. Just to name a few works, Girvan and Newman (2002) proposed a method to find communities using the idea of centrality indices to delimit the contour of the community. For complex networks, Capocci *et al*. (2005) developed an algorithm to detect communities based on spectral methods and take into account the weight at the edges and the orientation of the connections. Along the same lines, Clauset, Newman and Moore (2004) present a hierarchical agglomeration algorithm that aims to be fast in the detection of communities in large networks.

The main objective of community detection is to separate a network into groups of vertices with few connections between them. A measure used for this purpose is modularity, which identifies the contrast in edge density within the group, compared to the expected value for a random distribution of edges, that is, it measures the quality of each partition.

An expression to calculate a number that expresses modularity in a network can be found, for example, in Barabási (2016), and is given by:

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

In the expression, $n_c$ is the number of communities, $L_c$ is the number of edges in the community, $L$ is the total number of edges and $k_c$ is the total number of degrees of the vertices in the community. The higher the value of $M$ (which does not exceed 1), the better the partition of the community structure. The value of $M = 0$ expresses that the total network is a simple community. Negative modularity occurs if each vertex belongs to separate communities. To show communities of words in the networks of titles of the master's dissertations studied in this work, we opted for the Louvain Method (Blondel *et al*., 2008), which has a fast and good precision algorithm and the detection of communities is based on the measure of modularity. The algorithm implemented in Pajek (Batagelj and Mrvar, 1998) introduces a resolution parameter r, which converts Equation 1 into:

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

# 3. METHODOLOGY

## 3.1 DATA COLLECTION

The preliminary stage of data collection consisted of listing the courses recommended by CAPES (Coordination of Superior Level Staff Improvement) linked to the Ministry of Education of Brazil (MEC), after consulting the official website http://www.capes.gov.br/ (accessed September 2013). The search was limited to the large area of "Exact and Earth Sciences" in the assessment area of "Mathematics (Mathematics / Probability and Statistics)". Of the 45 courses found on the site, the 42 selected courses were specific to math. In the next step, it was discovered that in one of the courses there were no defended dissertations yet, so the final number of courses was 41, as shown in Table 1.

| n. | Course name | IHE | FU |
|---|---|---|---|
| 1 | Mathematics | FUFPI | PI |
| 2 | Mathematics | IMPA | RJ |
| 3 | Mathematics | PUC-RIO | RJ |
| 4 | Applied and Computational Mathematics | UEL | PR |
| 5 | Mathematics | UEM | PR |
| 6 | Mathematics | UFABC | SP |
| 7 | Mathematics | UFAL | AL |
| 8 | Mathematics | UFAM | AM |
| 9 | Mathematics | UFBA | BA |
| 10 | Mathematics | UFC | CE |
| 11 | Mathematics | UFCG | PB |
| 12 | Mathematics | UFES | ES |
| 13 | Mathematics | UFF | RJ |
| 14 | Mathematics | UFG | GO |
| 15 | Mathematics | UFJF | MG |
| 16 | Mathematics | UFMA | MA |
| 17 | Mathematics | UFMG | MG |
| 18 | Mathematics and Statistics | UFPA | PA |
| 19 | Mathematics | UFPB_JP | PB |
| 20 | Mathematics | UFPE | PE |
| 21 | Applied Mathematics | UFPR | PR |
| 22 | Mathematics | UFRGS (ma) | RS |
| 23 | Applied Mathematics | UFRGS (maplic) | RS |
| 24 | Mathematics | UFRJ (ma) | RJ |
| 25 | Applied Mathematics | UFRJ (maplic) | RJ |
| 26 | Pure and Applied Mathematics | UFSC | SC |
| 27 | Mathematics | UFSCar | SP |
| 28 | Mathematics | UFSM | RS |
| 29 | Mathematics | UFU | MG |
| 30 | Mathematics | UFV | MG |
| 31 | Mathematics | UNB | DF |

|  |  |  |  |
|---|---|---|---|
| 32 | Applied and Computational Mathematics | UNESP_PP | SP |
| 33 | University Mathematics | UNESP_ RC | SP |
| 34 | Mathematics | UNESP_ SJRP (ma) | SP |
| 35 | Applied Mathematics | UNESP_ SJRP (maplic) | SP |
| 36 | Mathematics | UNICAMP (ma) | SP |
| 37 | Applied and Computational Mathematics | UNICAMP (macomp) | SP |
| 38 | Applied Mathematics | UNICAMP (maplic) | SP |
| 39 | Mathematics | USP (ma) | SP |
| 40 | Applied Mathematics | USP (maplic) | SP |
| 41 | Mathematics | USP/SC | SP |

**Table 1.** List of courses used in the research.

Table with the name of the course, Institution of Higher Education (IHE) to which it is linked and Federative Unit (FU) of the Institution unit.

For the construction of the networks, the main source for obtaining the data were the CAPES indicator books on the site: http://conteudoweb.capes.gov.br/conteudoweb/CadernoAvaliacaoServlet. On the site are the notebooks from 1998 to 2012, in which the assessments of the programs were carried out every three years and served as the basis for the compilation. With the change in the assessment by four-year period, the 2016 data were not considered. Once the period of interest and the desired institution were chosen, the area to be analyzed was investigated, in this case, Mathematics / Probability and Statistics. The group entitled "*TE-Teses e Dissertações*" (Thesis and Dissertations) when activated shows a PDF document with several delimited fields, from which [author], [title] and [line of research] were obtained. Data from all fields was transferred to an organized spreadsheet with one field for each column. The procedure was successful for each course from 1998 to 2012.

## 3.2 NETWORK CONSTRUCTION

The semantic networks of titles of the master's dissertations were constructed from the data sheet. Each title is copied on a line in a text file that is subject to two types of treatment: the first is to eliminate punctuation marks such as commas, periods, hyphens and other signs. The second is more complex and uses a routine developed by Caldeira (2005) that uses three programs: (i) the UNITEX package, available at http://www-igm.univ-mlv.fr/unitex/; (ii) the Ambisin program, developed by Caldeira (2005) to address issues such as the elimination of ambiguities, the elimination of grammatical words and the separation of flexed or canonical word forms; (iii) the NetPal program, developed by Professor Dr. José Garcia Vivas Miranda and his collaborators, which generates the network in the appropriate format for the use of the Pajek program, created by Vladimir Batagelj and Andrej Mrvar (Batagelj and Mrvar, 1998).

After the construction of the networks for each course, the files were grouped according to the geographical region in which the headquarters of each program or course is settled. This division allows to investigate regional characteristics in the election of titles of the master's dissertations.

## 3.3. APPLICATION OF THE LOUVAIN METHOD

The algorithm consists of optimizing the modularity to find the partition that results in a higher modularity value. The full description of the algorithm, summarized here, is found in Blondel *et al*. (2008). The algorithm is based on two steps that must be iteratively performed. In the first step, all vertices are considered to form their own community, that is, we have *N* vertices forming *N* communities. In the second step, an ordered search is performed on all vertices from 1 to *N*, so that the neighboring vertex is incorporated into that community if there is an increase in modularity. This step is performed iteratively until maximum local modularity is achieved. In this step, each vertex can be visited several times. Once the maximum location has been reached, the algorithm builds a new network in which the vertices are the communities found with the weights of the ties between the communities calculated as the sum of the total weights between the vertices of those communities. The second step, as noted, is repeated iteratively (on the network whose vertices represent communities), which leads to a hierarchical decomposition of the network. This makes the algorithm suitable for handling large networks.

### 3.3.1. IMPLEMENTATION OF THE PAJEK ALGORITHM

There is a description of how to use the Louvain method implemented in Pajek on the Andrej Mrvar website, http://mrvar.fdv.uni-lj.si/pajek/community/Community-DrawExample.htm. Without going into technical details inherent to the program, only some pieces of information we consider relevant are highlighted in this study:

1. There are two ways to access different routines of the algorithm. The first, called Multi-LevelCoarsening + Single Refinement, only refines the partition obtained at the last level (the least refined partition). The second, called Multi-LevelCoarsening + Multi-LevelRefinment, differs from the first in that it makes coarse and refined partitions for each level obtained.

2. It is recommended to test the algorithm for different values of the resolution parameter *r* (Equation 2), which by default value is 1. High resolution values produce a large number of communities, while low values (greater than 0) produce few communities.

3. For best results, even without maximizing modularity, it is suggested to compare the partitions obtained in two rounds of algorithm execution with the same resolution parameter to assess the correlation. Pajek program has routines for this evaluation: Cramer's V, Rajski and the adjusted index of Rand. If the correlation between the two partitions is small, the number of communities is probably not correct and, therefore, the algorithm must be run with another value (greater or lesser) of the resolution parameter *r*. It is also suggested to use the highest correlation index found, even if the modularity is not the greatest.

These procedures were used for the semantic networks of titles of the master’s dissertations, separated by geographical regions, as well as for the total network, that is, for the network with all the titles of the master’s dissertations that form the largest connected component.

## 4. RESULTS AND DISCUSSION

### 4.1 NETWORKS BY COURSE

Table 2 shows the main quantities found to characterize the semantic networks of titles of the master’s dissertations for each course.

As can be deduced from Table 2, approximately half of the courses (21 courses) produced dissertations during the entire collection period. Of the remaining 20, two courses of Applied Mathematics (from UFRJ and UNESP-SJRP) were closed before 2012 and 18 courses began after

| IHE | Period | *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 |

**Table 2.** Quantities for semantic networks of titles of the master's dissertations per course.

Note: *QT*- number of titles; *QV* - number of vertices; *QA(p* = 1) - number of edges with weight 1; *QA*(*p*> 1): number of edges with a weight greater than 1; % *MC*: percentage of the largest component (in terms of number of vertices); *DR*-reference diameter.

(*) The network with a reference diameter of 1 is a unique case of SNT-UFMA that has only 4 titles, 3 of which form the largest component as a clique.

1998. These amounts reflect the growth in the offer of Masters in Mathematics in the selected period.

The number of titles of the master's dissertations varies between 4 and 205 titles, that is, there is a great variation in these amounts, mainly influenced by the period of operation of the course. For example, the Master in Mathematics course at UFMA presents only 4 titles of the master's dissertations because this course was implemented in 2012, the final collection date of three-year period. On the other hand, the variation in the number of titles of the master's dissertations for courses in operation from the beginning (1998) to the end (2012) of the collection, shows the difference in production in them.

The size of each network given by the number of vertices shows that from the point of view of complex networks, they are not large networks, the largest network has 618 vertices (UFRGS - Applied Mathematics). The average size of the titles of the master's dissertations, which is the ratio between the number of vertices (*n*) and the number of titles of the master's dissertations ($n_q$), varies between 2.26 and 7.43. This ratio reveals the greater or lesser diversity of words in the choice of titles of the master's dissertations. Comparing SNT-IMPA that has 51 titles and 253 different words with SNT-UFC that has 163 titles and 369 different words, the analysis of the $n/n_q$ ratio (SNT-IMPA: 4.61; SNT-UFC: 2.26) shows that a program may have fewer titles of the master's dissertations than another, but proportionally, its vocabulary is more diverse in comparison to the number of titles used.

Table 2 also provides information on the number of connections between the vertices, that is, the number of connections established between the words of the titles of the master's dissertations for each course. As the number of times a link between two words is considered, columns 3 and 4 differentiate the links that occurred only once (weight 1) from the links that occurred more than once (weight greater than 1). The results show the predominance of connections with weight 1, that is, most connections between words occur only once.

The number of components is one of the parameters that indicates how fragmented the network is, that is, how many groups of words are not linked. It can be seen in Table 2 that of the 41 networks, only 5 of them (approximately 12%) have only 1 component. These data would indicate the greater or lesser diversity in the choice of word groups to compose the titles of the master's dissertations. However, the percentages of the largest components, which indicate the number of words of these in relation to the total network, show that the smaller components are not significant for this interpretation. In fact, it is observed that the size of the largest component is above 90% of the vertices in 29 courses (70.7%). It is also observed in the remaining networks, that is, with the largest component with less than 90% of vertices, all are incomplete with respect to the period collected (1998 to 2012). Therefore, there is a tendency for this percentage to grow. Figure 1 shows the relationship between the number of titles of the master's dissertations and the percentage size of the largest component of the SNT.

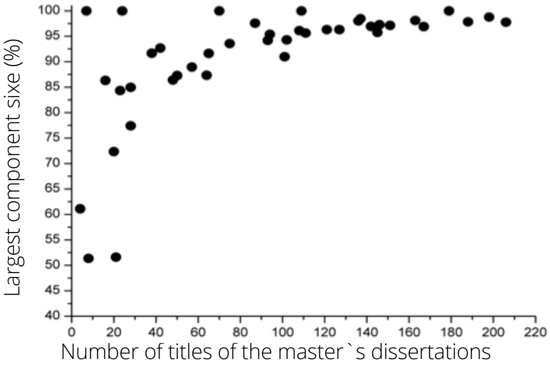

**Figure 1.** Graph of the size distribution of the largest component in the number of titles of the master's dissertations.

Figure 1 shows that networks with a greater number of titles of the master's dissertations tend to exhibit larger components with sizes close to 100%. Therefore, if two networks with comparable titles amounts show a significant difference in the size of the largest component, this can be attributed to the choice of different topics in the SNT with a lower percentage of the component. For example, SNT-UEM (101 titles of the master's dissertations; largest component 86.32%), compared to

SNT-UFScar (102 titles of the master's dissertations; largest component 94.29%) can show that it presents more diverse themes. This indication is qualitatively associated with the fact that the UEM program is highlighted in three areas (algebra, analysis and geometry) while the UFScar program is highlighted in only two areas (analysis and geometry).

Another network parameter that can be used to characterize programs or courses is the diameter of the network. The diameter of a connected network is the longest of all the calculated geodesic distance in a network. In the case of connected SNT, it represents the maximum space between two words present in different titles of the master's dissertations. Therefore, it refers to the diversity of topics used in the formation of titles of the master's dissertations. As Fadigas and Pereira (2013) define, the reference diameter is a normalization of the diameter relative to the largest possible diameter in a minimally connected clique structure, that is, each clique is connected to another by a single vertex. This structure is called "line structure". For SNT, Table 2 shows that only 1% of these are in the lowest range of the reference diameter (0 to 0.25), which corresponds to a "star structure" and, therefore, to a more cohesive network in the approximation between their titles of the master's dissertations. However, the reference diameter is influenced by the size of the network, that is, by the number of titles. To make comparisons possible, Figure 2 shows the variation of the reference diameter with the size of the titles of the master's dissertations for the SNT. According to Figure 2, there is a tendency to reduce the reference diameter with the increase in the number of titles. Therefore, the larger values of the reference diameter are more influenced by the smaller number of titles than by a possible diversity of vocabulary, for example.

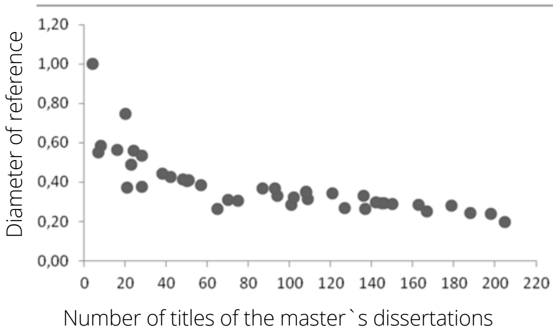

**Figure 2.** Graph of the variation of the reference diameter with the number of titles of the master's dissertations for the SNT.

## 4.2. COURSE NETWORKS GROUPED BY GEOGRAPHICAL REGION

### 4.2.1. GENERAL QUANTITATIVE DATA

We grouped the networks by geographic region of Brazil to have both a quantitative and qualitative macro approach to the semantic networks of titles of the master's dissertations in Mathematics. This allowed us to analyze the differences and similarities between the networks and infer some regional characteristics. For example, the information in Table 2 allows us to infer that the programs located in the Southeast Region were those that produced the most dissertations (1826), followed by the Northeast Region (736), the South.

Region (609), the Midwest Region (344) and the Northern Region (166). However, it is interesting to note that the average number of dissertations per region, when calculated by the number of courses, presents the Midwest Region as the one that produces the most (172 per course), while the other regions have an average between 83 and 92 dissertations per course, with the North Region as the least productive (83 per course) and tied with the Southeast Region. When the calculation takes into account the periods of each course, the Southeast Region presents a small increase in relation to the North Region (7.97 vs. 7.90), while the other regions continue in the same order of production. The fact that the Midwest region presents itself as the most productive is strongly linked to the production of UnB (Table 3).

The number of vertices in the network indicates the number of different words used in the set of titles of the master's dissertations, that is, excluding repeated words from that amount. To quantify word repetitions when the networks of each program come together to form the networks by region, we present an index that quantifies the reduction of common vertices. This index is calculated using Equation 3, in which S_(n_0 )  is the sum of the vertices of the network for each program and n is the total number of vertices for the network of programs in that region.

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

Table 3 shows some quantities for networks by region, in which it is observed that the network in the Southeast Region is the one with the highest *IRV*. The value (55.2%) indicates that more than half of the words used in the titles of the master's dissertations form a vocabulary from a source common to all courses in that region. When the number of titles is taken into account, the parameterization indicated by the *IRVp* shows that the SNT of the Southeast Region has the lowest index, while the others have indices in the same order of magnitude. This means that there is a greater "disciplinarity" in terms of vocabulary chosen for titles of

| Region Network | *NCour-se* | *QT* | *NComp* | *%MC* | *n* | *IRV* | *IRVp* |
|---|---|---|---|---|---|---|---|
| Midwest | 2 | 344 | 4 | 98.991 | 793 | 0.180 | 0.053 |
| Northeast | 8 | 736 | 3 | 99.674 | 1228 | 0.398 | 0.054 |
| North | 2 | 166 | 2 | 99.457 | 552 | 0.086 | 0.052 |
| Southeast | 22 | 1826 | 14 | 98.492 | 1295 | 0.552 | 0.030 |
| South | 7 | 609 | 5 | 98.996 | 2569 | 0.337 | 0.055 |

**Table 3:** Network characterization with emphasis on components.

Note: *NCourse* - number of courses; *QT*- number of titles; *NComp*- number of components; % *MC* -percentage of the largest component (in terms of number of vertices); *n* - number of vertices of the network; *IRV*- vertices reduction index; *IRVp* - parameterized vertex reduction index (*IRV / QT*).

the master's dissertations in the Southeast Region.

## 4.2.2. COMPONENTS AND COMMUNITY OF WORD

Information on the number of components of each network per region, combined with the percentage of vertices of the largest component (Table 3), indicates the behavior in relation to groups of words closer to those separated from other groups of words. Therefore, the presence of more than one component in the network means that there are groups of isolated words, for some reason that is peculiar to each network. As shown in Table 3, the largest component size for all networks is around 99%. This shows the predominance of the largest component over the smallest, that is, for all networks there is a predominance of a large group of words linked together. Therefore, there is no considerable difference in the topology of the networks in terms of the number of components and the size distribution of the components, indicating a uniformity in the choice of vocabulary used in the titles of the master's dissertations in each group of courses by region.

Although the analysis of the distribution of quantity and size of the components to provide a first approximation in terms of groups, it does not provide details about the characteristics of the largest component. The purpose of applying the algorithm to detect communities in semantic networks of titles of the master's dissertations by geographic region is to detect similarities and differences between the largest component of the networks and infer interpretations for the results. Word communities are characterized by having a greater number of internal links than expected for a random distribution of edges in the network.

The networks were built only with the vertices of the largest connected component for a more precise analysis of the formation of word communities, which in this case is representative of the networks as a whole considering the percentages of the largest components, according to Table 3. The algorithm of the Louvain Method (Blondel *et al*., 2008) is incorporated into the Pajek program (Batagelj eMrvar, 1998), used for community research. Its application, for more reliable results, requires the investigation of the optimal resolution r, a parameter that directly affects the modularity value and the number of communities: lower values result in a smaller number of communities.

In turn, modularity indicates the quality of the division of communities, high values (maximum

| Networks by Region | *R* | *M_r* | Cramer's V Index | Communities | *%MCommon* |
|---|---|---|---|---|---|
| Midwest | 0.25 | 0.751963 | 1.000000 | 3 | 96.56 |
| Northeast | 0.25 | 0.750069 | 1.000000 | 2 | 99.67 |
| North | 0.25 | 0.815102 | 0.974503 | 3 | 45.72 |
| Southeast | 0.10 | 0.900062 | 1.000000 | 2 | 98.89 |
| South | 1.50 | 0.399884 | 0.759705 | 19 | 9.91 |

**Table 4:** Quantities related to community detection.

1) imply dense connections between the vertices within the group and sparse connections between the vertices of different groups. However, modularity alone does not result in a precise determination of the number of groups in the network, since the same resolution can result in different modularities and group quantities when the algorithm is performed again. With respect to the implementation in Pajek and the application in the networks, tests were carried out to optimize the application of the algorithm following the guidelines described in subsection 2.3. For each network analyzed, the resolution values of 0.1; 0.2; 0.25; 0.5; 1.0; 1.25; 1.5 and 1.75 were used and the resolution that resulted in a higher value of the statistical index known as Cramer's V, which measures the correlation in the calculation of modularities for pairs of equal resolutions, was chosen. Columns 2 to 4 of Table 4 show the results for networks by region.

It can also be seen in Table 4 that, from the point of view of groups or communities of words, the networks can be divided into three categories: the first category includes the networks of the Midwest, Northeast and Southeast with a low number of communities (2 or 3) and a community with almost all vertices (more than 96.5%). In the second group is the network of the North region with only 3 communities but with the size of the communities well distributed (the other two smaller communities have percentages of 16.39% and 37.89%). Finally, the South Region network is in the third category with 19 communities with well distributed sizes between 1.79% and 9.91% (largest community included). It is worth mentioning that the last network is not very stable, considering resolution 1.5, resulted in a higher value of the Cramer's V index and the number of communities can still vary. For the other networks, the same number of communities was found in the different executions of the algorithm for optimal resolution.

From the point of view of the words used in the titles of the master's dissertations, the networks of the northeast, southeast and midwest regions behave as a cohesive group in which the connections between the words within a community are greater than expected, if the distribution was random. Modularity values between 0.750 and 0.900 confirm this behavior. In the case of the northern region network, this characteristic can also be inferred but the internal cohesion indicated by the modularity of 0.815 implies the distinction of three groups of words due to a certain homogeneity in the group sizes. The network of the southern region has the lowest modularity value (0.395), which indicates a weaker internal cohesion of the groups, that is, the connections are close to those expected for a random distribution of connections between words. From the point of view of modularity, the word network of the southeast region is the one with the best quality of community division, while the network of the southern region is the one with the weakest division.

There is a contrast taking into account the number of courses for these last two regions: the 22 courses in the southeast region are more homogeneous in terms of choosing the words of the titles of the master's dissertations compared to the 7 courses in the south region that are more heterogeneous.

## 5. CONCLUDING REMARKS

The analysis made from the semantic networks of titles of the master's dissertations for each of the programs/courses reveals that there is a diversity of operational periods, number of titles, number of vertices, amount of weighted edge, percentage of the largest component and reference diameter. However, such diversity does not prevent inferences about the long-term behavior of networks such as the asymptotic growth of the percentage of the largest component (tends to 100%) with the increase in the amount of titles of the master's dissertations and, therefore, in the amount of words (vertices) that form them.

The behavior of the reference diameter with the increase in the number of titles of the master's dissertations shows that there is a trend in all networks that, as new titles are added, the maximum geodetic distance between two words in the network decreases.

When the comparison is made with the SNT grouped by geographic region, it is observed that there is also a differentiation in the aspect of academic production, reflected by the number of titles of the master's dissertations in the regions and with emphasis on the southeast region, in absolute terms. However, the fact that the midwest region has only two institutions surveyed, including a remarkably productive institution like UnB, highlights it in terms of average productivity. In terms of the vocabulary used in the titles, the network of titles of the master's dissertations in the southeast region is less diversified, indicating greater "disciplinarity". The application of the Louvain method to detect communities proved to be appropriate to differentiate networks with greater or lesser diversity of cohesion between groups of words that formed the titles. The behavior of the south region network, which has a greater diversity of cohesion, is expressed by the greater number of communities and the lower modularity.

The use of Louvain's algorithm to detect communities proved satisfactory when used in semantic networks of titles of the master's dissertations, since it can identify and quantify the groups of words with greater internal cohesion, whose interpretation is initially linked to the discipline. This method can be explored together with other quantitative and qualitative analyzes that take into account the research lines of each course/program associated with the areas defined by the organizations that promote scientific and technological research such as the National Council for Scientific and Technological Development (CNPq) in Brazil, in order to identify issues of regional or national interest; if these are in line with what is being studied in the world and if there are focal points of characteristic studies in the country that justify more or less specific vocabularies.

# BIBLIOGRAPHY

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

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

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and 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., and 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., and 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., and 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., and 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., and 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. and 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. and 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., and 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., and 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., and 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., ... and 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.