

---

Edital Nº 02/2016 – PDPD

**Título do projeto:** Análise webométrica da UFABC: Uma abordagem computacional baseada em grafos de hiperlinks.

**Nome do aluno:** Lara Tenore Ferreira

**RA do aluno:** 11030616

**E-mail do aluno:** [lara.tenore@aluno.ufabc.edu.br](mailto:lara.tenore@aluno.ufabc.edu.br)

**Nome do orientador:** Jesús P. Mena-Chalco

**E-mail do orientador:** [jesus.mena@ufabc.edu.br](mailto:jesus.mena@ufabc.edu.br)

**Palavras-chave:** Análise de dados, webometria, website da UFABC, hiperlinks.

**Área de conhecimento do projeto:** Ciência da computação

---

23 de setembro de 2017

**Universidade Federal do ABC  
Santo André, São Paulo**

À Ilustríssima Pró-Reitora de Pesquisa, Prof<sup>a</sup>. Dr<sup>a</sup>. Marcela Sorelli Carneiro Ramos,

Encaminho o relatório final da aluna Lara Tenore Ferreira referente ao projeto de pesquisa junto ao programa PDPD no edital Nº 02/2016.

É importante destacar que a Lara: (1) apresentou ótima autonomia acadêmica, e (2) teve um excelente desempenho no desenvolvimento das atividades de estudo/aprendizado dos algoritmos considerados no Projeto de Iniciação Científica.

Atenciosamente,



Jesús Pascual Mena-Chalco  
**Professor Adjunto - CMCC**  
SIAPE 1934625  
jesus.mena@ufabc.edu.br

# Análise webométrica da UFABC: Uma abordagem computacional baseada em grafos de hiperlinks

Lara Tenore Ferreira  
RA 11030616  
Bacharelado em Ciência & Tecnologia  
Universidade Federal do ABC  
[lara.tenore@aluno.ufabc.edu.br](mailto:lara.tenore@aluno.ufabc.edu.br)

Orientador:  
Jesús P. Mena-Chalco  
[jesus.mena@ufabc.edu.br](mailto:jesus.mena@ufabc.edu.br)

25 de setembro de 2017

## Resumo

A webometria é entendida como a aplicação de métodos computacionais para a medição quantitativa de informações (e.g., indicadores) registradas em páginas web tais como tamanho médio de uma página, visibilidade (importância na web), densidade, fator de impacto. A webometria permite estudar e caracterizar o espaço web virtual de determinados atores, grupos ou instituições. Neste projeto de Pesquisa desde o Primeiro Dia (PDPD) propõe-se a caracterização do espaço web endógeno (páginas web internas) da Universidade Federal do ABC (UFABC), dividindo-o conforme sub-domínios web.

O método desenvolvido considerou os seguintes três grandes processos: (i) Criação do grafo de hiperlinks, (ii) Calculo de informações topológicas, como por exemplo, densidade, grau de entrada e grau de saída, e (iii) Análise dos dados considerando estatística descritiva, coeficientes de correlação e número de páginas externas e número de arquivos. Acreditamos que a coleta de informações e sua análise permitiu revelar características importantes da UFABC sobre a ótica webométrica.

**Palavras-chave:** análise de dados, webometria, website da UFABC, hiperlinks.

## Sumário

<b>1</b>	<b>Introdução</b>	<b>5</b>
1.1	Objetivo Geral . . . . .	5
1.2	Objetivos Específicos . . . . .	5
1.3	Metas . . . . .	6
<b>2</b>	<b>Webometria</b>	<b>6</b>
2.1	Indicadores e ferramentas webométricas . . . . .	7
2.2	Dificuldades da Webometria . . . . .	8
<b>3</b>	<b>Método desenvolvido na Iniciação Científica (PDPC)</b>	<b>8</b>
3.1	Criação do grafo de hiperlinks . . . . .	9
3.2	Cálculo de informações básicas . . . . .	11
3.3	Análise dos dados . . . . .	11
<b>4</b>	<b>Cronograma</b>	<b>12</b>
<b>5</b>	<b>Resultados</b>	<b>13</b>
<b>6</b>	<b>Conclusões</b>	<b>20</b>
<b>A</b>	<b>Algoritmo para criação do grafo de hiperlinks</b>	<b>21</b>
<b>B</b>	<b>Grafos de hiperlinks do <i>website</i> da UFABC</b>	<b>24</b>

# 1 Introdução

Conhecer as características da própria instituição acadêmica sempre foi algo importante, tanto para saber o caminho que ela está seguindo como que rumo tomará. Possuir formas de obter essas informações é um problema ainda mais frequente devido à grande quantidade de dados e informações gerados diariamente, principalmente dados na internet, considerando que é parte da cultura acadêmica compartilhar informações (Thelwall, 2009).

Uma das possíveis técnicas utilizadas para resolver esse problema é a webometria que permite analisar de forma quantitativa o conteúdo web de, por exemplo, instituições e, dessa maneira, fornecer uma possível aproximação entre a tecnologia e a sociedade, de tal forma que essas instituições possam tomar decisões mais precisas acerca de diversos assuntos baseadas nos estudos webométricos.

A webometria vem sendo amplamente estudada nas áreas de Ciência da Computação, Ciometria e Ciência da Informação e áreas correlatas, uma vez que aprender e compreender a construção no espaço web e a estrutura de grafos são essenciais para entender a presença, influência e evolução de determinada empresa, instituição ou país na web de maneira clara e, de certa forma, exata.

Por outro lado, a Universidade Federal do ABC é conhecida pelo seu projeto inovador e interdisciplinar. Em agosto de 2013 a UFABC alcançou as primeiras colocações entre instituições nacionais referentes a qualidade e impacto da sua produção científica no ranking Scimago-2013.

Considerando o espaço web: Quais as principais características topológicas presentes no grafo de hiperlinks mantida pela universidade? Quais são as páginas web mais importantes da instituição? Nesse estudo, procuramos responder essas perguntas e compreender a dinâmica apresentada nas páginas web da universidade.

Este projeto poderá auxiliar na análise da páginas web institucional da UFABC, analisando diferentes aspectos da mesma por técnicas webométricas com o intuito de identificar padrões e levantar possibilidades para uma futura melhora da informações web da nossa universidade.

### 1.1 Objetivo Geral

O objetivo geral deste projeto PDPD é a análise da rede de hiperlinks construído a partir do website da UFABC usando conceitos de webometria.

### 1.2 Objetivos Específicos

- Estudo de técnicas computacionais para a construção de uma rede complexa de hiperlinks.
- Estudo de formas de caracterização do grafo utilizando webometria.

Além dos objetivos específicos de pesquisa, o projeto também tem os objetivos característicos de um projeto de Iniciação Científica padrão:

- Obter resultados inéditos e relevantes para os problemas estudados, o que contribui para o próximo objetivo;

- Incentivar o gosto da aluna pela pesquisa e a criatividade para solução de problemas;
- Envolver a aluna em problemas de pesquisa na área de Ciência da Computação;
- Desenvolver a maturidade algorítmica da aluna;
- Ampliar os conhecimentos e habilidades da aluna, especialmente aqueles relacionados aos assuntos de Ciência da Computação e também aqueles relacionados ao desenvolvimento de pesquisa em geral, tais como escrita de artigos e relatórios, preparação e apresentação de seminários, etc;
- Desenvolver na aluna prática de leitura e compreensão de artigos científicos, através do estudo de artigos especialmente selecionados;
- Ampliar os conhecimentos da aluna na língua inglesa. Na área da Ciência da Computação, a maioria dos artigos e livros são publicados somente em língua inglesa e a escrita em inglês é importante na produção de textos para submissão em conferências e revistas.

### 1.3 Metas

- Construção da rede de hiperlinks de páginas web da UFABC.
- Visualização da rede de hiperlinks.
- Criação de material didático sobre webometria.

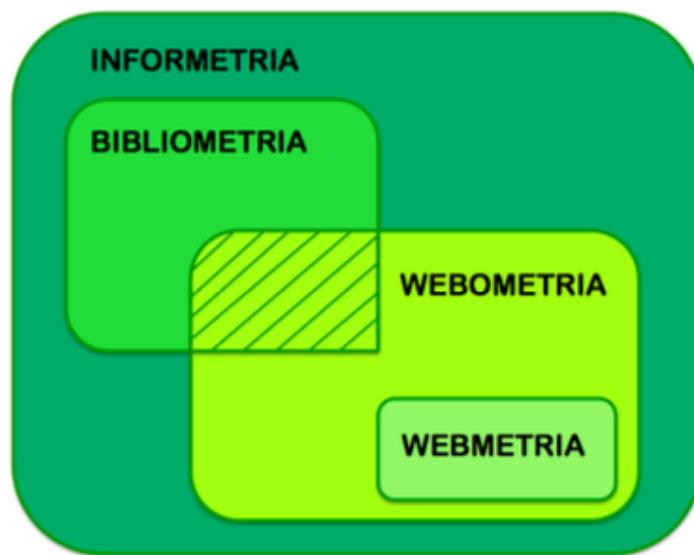
## 2 Webometria

Etimologicamente, a palavra webometria é a junção da palavra Web, que é um sistema de informações por hipermídia que permite que um usuário tenha acesso a diferentes conteúdos na internet, e do sufixo “metria”, que indica o ato de medir. Sendo assim, webometria é uma área que analisa quantitativa e, consequentemente, qualitativamente a World Wide Web e suas páginas específicas, com um intuito de fornecer para, desde pequenas empresas até grandes instituições e países, um estudo sobre sua respectiva presença e evolução no espaço da web ([Vanti, 2002](#)). A webometria permite uma aproximação entre o tecnológico e o social, possibilitando que decisões possam ser tomadas a partir dessas análises ([Thelwall, 2012](#)).

O termo webometria foi utilizado pela primeira vez por Tomas C. Almind e Peter Ingwersen em 1997, mas encontra-se também na literatura como cybermetria e internetometria por diferentes autores ([Vanti, 2002](#)). Neste projeto de pesquisa optaremos apenas pelo primeiro termo por ser o mais utilizado na maioria das referencias existentes sobre o tema.

Vale ressaltar também que o termo webometria não pode ser confundido com o termo webmetria, sendo este um sub-campo do primeiro que aborda um tipo de análise específico de relatório de dados. Veja na Figura 1 a relação entre a informetria, bibliometria, webometria e webmetria.

Para a realização das análises, a webometria utiliza-se de métodos e abordagens bibliométricas e, principalmente, informétricas, bem como suas próprias ferramentas, instrumentos de medida e classificações de páginas da web (páginas web pessoais, institucionais ou organizacionais, ad hoc ou páginas).



**Figura 1:** Diagrama da inter-relação entre informetria, bibliometria, webometria e webmetria. Imagem adaptada do trabalho de [Vanti \(2002\)](#).

## 2.1 Indicadores e ferramentas webométricas

A webometria possui algumas ferramentas e índices que facilitam e executam o estudo na WWW. São eles:

- Tamanho: mede a quantidade de páginas que o site possui ([da Silva, 2011; Vanti, 2007](#)).
- Fator de impacto na Web (*FIW*) – Segundo Mike Thelwall, consiste no número de páginas que levam a um determinado site ou área da internet, dividido pelo número de páginas neste site ou área. O *FIW* pode ser expressado pela seguinte formula:

$$FIW = \frac{\text{Número de páginas que referenciam determinado site}}{\text{Número de páginas do site referenciado}}$$

Este indicador permite estudar a presença e grau de reconhecimento de um site ou de um país no espaço da Web. Devido à natureza dinâmica e em tempo real da rede, esse indicador permite analisar esses fatores de um site em um determinado ponto no tempo, podendo assim, complementar os resultados aos obtidos com as medições tradicionais ([Vanti, 2002](#)).

- Luminosidade: mede o número de links externos presentes no site, o que determina como o site analisado se conecta com o resto da web (Vanti, 2007).
- Densidade da rede: relação que pode ser estabelecida entre o tamanho de uma página e número de links que ela possui. Quanto menor for o tamanho da página, mantendo o número de links, menor será a densidade. Quanto maior for a densidade de uma página, esta será mais descritiva e auto-suficiente (da Silva, 2011; Vanti, 2007). A densidade de rede pode ser expressada pela seguinte fórmula:

$$DR = \frac{\text{Número de links}}{\text{Número de páginas} \times (\text{Número de páginas} - 1)}$$

Os estudos webométricos, em sua maioria, são feitos com instrumentos proprietários como os motores de busca como por exemplo Google, Alta Vista, Yahoo, Hotbot. Esses instrumentos facilitam na quantificação, na avaliação dos fluxos de informação e no intercâmbio de dados na Web (Jeyashree and Ravichandran, 2013).

Segundo Alastair Smith, esses buscadores permitem contabilizar o número total de páginas na Web que se referem a certo assunto ou palavras chaves (Vanti, 2002), entretanto, não possibilitam a exploração dos dados por método mais sofisticados como os de grafos (ou redes). Neste projeto usamos os links das páginas web para criarmos um grafo de hiperlinks da UFABC.

### 2.2 Dificuldades da Webometria

No campo da webometria existem algumas dificuldades encontradas por pesquisadores. Facilmente links podem surgir e desaparecer. Isso provoca uma inconsistência no trabalho dos motores de busca e no estudo em geral (Gouveia, 2012), fazendo com que a quantidade de páginas e links que remetem a uma site específico possam aumentar ou diminuir em pouco tempo.

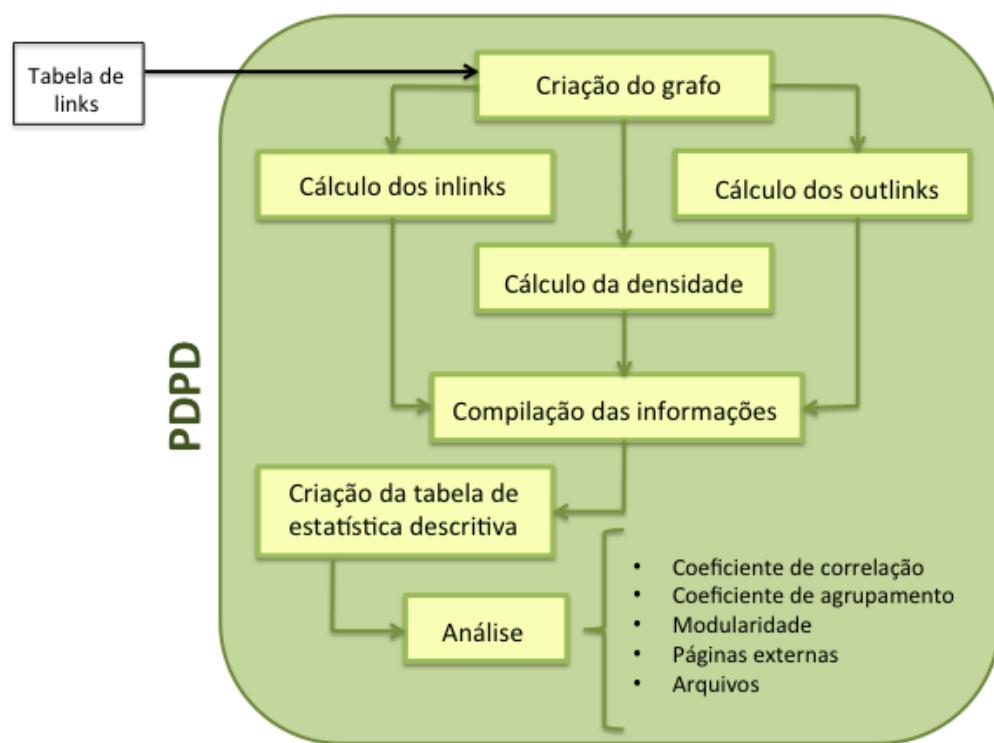
Quanto aos motores de busca, por serem muitos, é comum que seus resultados não se sobreponham e apresentem diferentes números para um mesmo tema ou palavra-chave. Além disso, por serem uma ferramenta automática, esses instrumentos muitas vezes não conseguem manter uma hierarquização e tendem a misturar dados de páginas web irrelevantes com os de páginas de maior importância no campo (Gouveia, 2012; Vanti, 2002).

Por fim, por mais que seja um campo emergente e cada vez mais presente na sociedade, muitas vezes a rede não reflete com total fidelidade a evolução ou o retrocesso de uma determinada empresa, país, assunto, tema ou disciplina. Ainda existem áreas que concentram suas informações em materiais impressos e, portanto, não podem ser analisadas por meios webométricos.

## 3 Método desenvolvido na Iniciação Científica (PDPD)

Este projeto teve como objetivo desenvolver e implementar um método que possibilitou analisar a página web institucional da UFABC. Na Figura 2 apresentamos um diagrama que re-

presenta o método adotado neste projeto de PDPD. Foram considerados três grandes processos: (i) Criação do grafo de hiperlinks, (ii) Cálculo de informações topológicas, como por exemplo, densidade, grau de entrada e grau de saída, e (iii) Análise dos dados considerando estatística descritiva, coeficientes de correlação, coeficiente de agrupamento, páginas externas e arquivos.



**Figura 2:** Fluxograma atualizado do método considerado para análise webométrica da UFABC. Em amarelo estão indicados os itens realizados neste PDPD. Fonte: Figura elaborada pelos autores do projeto.

### 3.1 Criação do grafo de hiperlinks

Nesse processo construímos um grafo de hiperlinks utilizando uma tabela de links previamente calculada considerando todas páginas web acessíveis a partir do site principal da UFABC.

Nesse sentido, foi desenvolvido um programa que tendo como entrada uma tabela de hiperlinks, gera um grafo que corresponde a todos hiperlinks internos. Nesse grafo, cada página representa um vértice, e cada hiperlink uma aresta direcionada.

O algoritmo desenvolvido permite identificar de forma única cada página (através de uma busca linear). As arestas identificadas correspondem a hiperlinks que estão dentro do domínio, isto é, apenas consideramos hiperlinks à páginas internas.

Os hiperlinks à páginas externas permitem evidenciar como as páginas da instituição referem-se a outras.

renciam fontes externas. Aqui é importante ressaltar que para cada vértice (i.e., página web interna) foram identificados valores quantitativos relacionados com o número de arquivos físicos disponíveis em cada página.

Nesse sentido, ao todo consideramos 11 tipos de arquivos que podem ser referenciados nas páginas web (Ver Tabela 1). Acreditamos que a informação sobre o tipo de arquivos referenciados em cada página permite também caracterizar a instituição sob a perspectiva da disponibilidade de informações.

**Tabela 1:** *Tipos de arquivos (extensões) considerados neste trabalho.*

<b>Tipo de Arquivo</b>	<b>Extensões</b>
<b>Documento textual</b>	doc, docx, rtf, odt.
<b>PDF</b>	pdf.
<b>Áudio</b>	mp3, wma, aac, wav, ac3.
<b>Vídeo</b>	mp4, avi, mpeg, mov, rmvb.
<b>Imagem</b>	jpg, jpeg, png, gif, tif.
<b>Apresentação</b>	ppt, pptx, odp.
<b>Planilha</b>	xls, xlsx, ods, odp.
<b>Hipertexto</b>	htm, html, oth, php.
<b>Compactado</b>	zip, rar, tgr, tar.gz, 7z, tar, xz.
<b>Executáveis</b>	exe, bin, sh.
<b>Arquivo de texto</b>	txt

### Algoritmo para criação do grafo de hiperlinks

Durante o projeto, foi desenvolvido um algoritmo para transformar a lista de páginas web de um domínio com suas respectivas URLs das páginas ao qual a página aponta. A lista de páginas web foi obtido no contexto de um trabalho de PDPD-2016 do aluno Lucas Theodoro Guimarães de Almeida<sup>1</sup>.

A transformação envolve a geração de um grafo, onde os vértices são páginas web e as arestas com conexões (*links*) entre elas. Além disso, o algoritmo é capaz de identificar quais os vértices estão relacionados a páginas externas e quais são os sites que representam arquivos online.

Os passos utilizados para a transformação são os seguintes:

- (a) Leitura do arquivo de entrada para identificar as páginas internas, representadas pelos vértices. Nesta etapa, são identificados também os arquivos em cada página para armazená-los ao vértice correspondente. Os arquivos são identificados através de um método chamado INDICEPAGINA, no qual, pelo final da URL, são separados de acordo com as extensões, especificadas na Tabela 1.
- (b) Segunda leitura do arquivo de entrada para identificar as conexões, representadas pelas aresta, de acordo com o vetor de vértices já identificados.

<sup>1</sup>O trabalho intitulado “Inter- e intra-relação dos sub-domínios da UFABC através de grafos de hiperlinks: Uma abordagem webométrica” foi concluído em setembro de 2017.

- (c) Geração do arquivo de saída no formato *Geographic Data File* (GDF) de acordo com os vértices, arquivos e arestas identificados nas fases anteriores.

No Apêndice A apresentamos o código-fonte do algoritmo descrito nesta seção. O programa foi desenvolvido na Linguagem de programação Python 3. A lista de páginas é um arquivo de texto com as URLs visitadas de forma sistemática (resultado da projeto do aluno Lucas Theodoro Guimarães de Almeida). O arquivo de saída é um arquivo em formato GDF que é utilizado para visualização no software Gephi (Bastian et al., 2009).

### 3.2 Cálculo de informações básicas

Nesse módulo foi realizada a caracterização webométrica do grafo construído anteriormente, utilizando análises webométricas quantitativas. Para cada grafo de hiperlinks gerado obtivemos medidas como grau médio, grau de entrada, grau de saída, densidade, modularidade e número de arquivos presentes em cada página. Para essa fase do processo utilizamos ferramentas especialistas para a análise dos grafos (e.g., gephi).

Os arquivos analisados foram: documentos de texto, PDF, audios, vídeos, imagens, apresentações, planilhas, arquivos de hipertextos, e arquivos compactados. Na Tabela 1 pode-se verificar de maneira mais detalhada quais extensões foram consideradas para tal análise.

### 3.3 Análise dos dados

Nesse módulo foram estudados e aplicados os conceitos básicos de estatística descritiva para determinar relações entre todas as características identificadas no processo anterior. Temos obtido informações importantes que permitem caracterizar instituições acadêmicas a partir de informações obtidas das suas páginas web (Orduña-Malea and Aguillo, 2015).

Aqui é importante destacar dois conceitos que são importantes para a análise dos dados: (i) coeficiente de correlação, e o (ii) coeficiente de agrupamento.

- **Coeficiente de correlação**

Segundo Figueiredo Filho and Silva Junior (2010), correlação é “uma medida de associação bivariada (força) do grau de relacionamento entre duas variáveis”. Adicionalmente, a correlação pode ser utilizada para mensurar a direção e o grau de relação entre duas variáveis quantitativas. Em suma, o coeficiente de correlação de Pearson é uma medida que associa variáveis de escala métrica, de maneira linear (Figueiredo Filho and Silva Junior, 2010). Este é representado pela seguinte fórmula:

$$r = \frac{N \sum_{i=1}^N XY - \sum_{i=1}^N X \sum_{i=1}^N Y}{\sqrt{N \sum_{i=1}^N X^2 - (\sum_{i=1}^N X)^2} \sqrt{N \sum_{i=1}^N Y^2 - (\sum_{i=1}^N Y)^2}}$$

O coeficiente pode assumir valores entre -1 e 1, sendo que 1 significa uma correlação perfeita entre as variáveis, -1 significa uma correlação imperfeita entre as variáveis (quando

uma aumenta, a outra diminui), e 0 significa que as duas variáveis independem linearmente uma da outra.

- **Coeficiente de agrupamento**

Esta medida define o grau com que os nós do grafo tendem a se agrupar em grupos coesos devido a um grande número de ligações entre os mesmos nós. O cálculo desta métrica leva em consideração à ligação entre três nós consecutivos e a formação de triângulos entre três nós. Assim, o coeficiente de agrupamento pode ser expressado da seguinte fórmula:

$$C = \frac{3 \times \text{número de triângulos}}{\text{Número de trio de vértices conectados}}$$

O coeficiente de agrupamento pode variar entre 0 e 1. Quanto mais próximo de 1, mais conectados estão os vértices.

## 4 Cronograma

Na Tabela 2 é apresentada a atribuição de cada atividade para o ano de estudo. As atividades desenvolvidas estão representadas com o símbolo ‘O’.

**Tabela 2:** Atividades realizadas no projeto.

	Out	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago
Revisão bibliográfica	O	O	O	O	O	O	O	O	O	O	O
Criação do grafo hiperlinks da UFABC	O	O	O			O	O				
Estudo da ferramenta de visualização de grafos e análise estatístico				O	O		O	O			
Calculo de informações topológicas básicas				O	O	O	O	O			
Análise dos dados								O	O	O	
Escrita dos relatórios de pesquisa e criação do material didático				O	O				O	O	

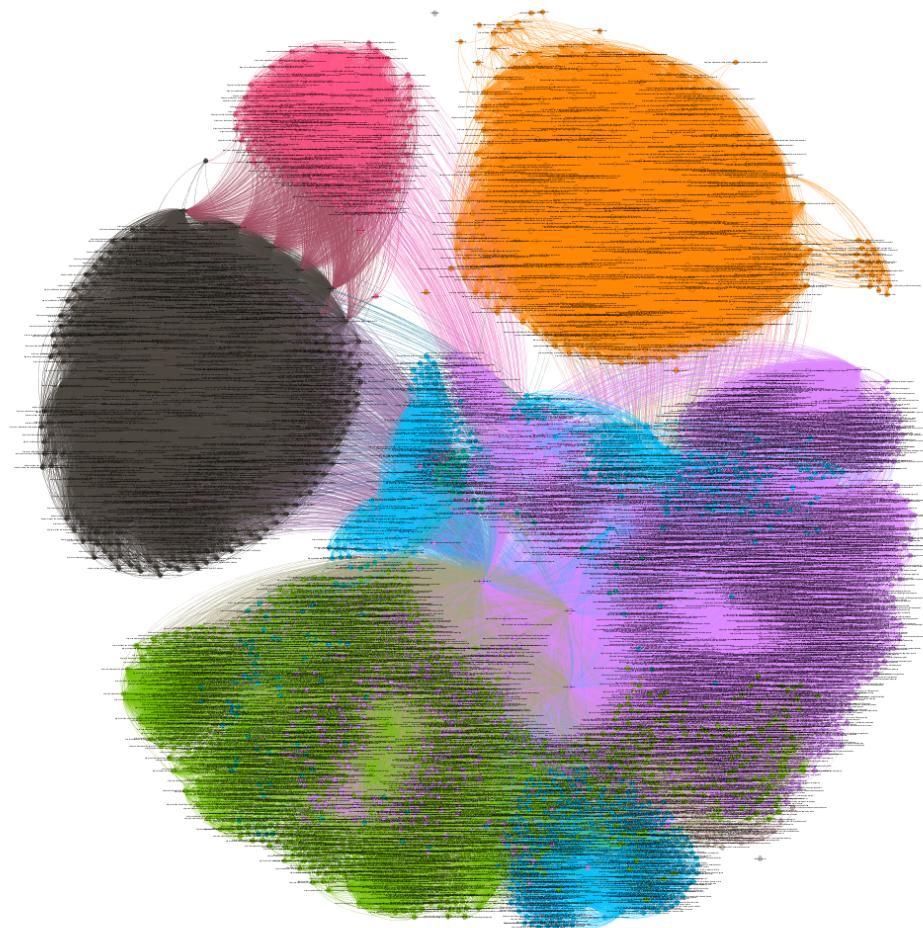
## 5 Resultados

Nesta seção apresentamos os novos resultados obtidos após o Relatório Parcial.

Na Figura 3 apresentamos o grafo de hiperlinks que é a composição dos grafos obtidos para os três centros da UFABC (CECS, CCNH e CMCC). Em laranja, temos as páginas CECS. Em preto e rosa, as páginas do CMCC. Em azul, verde, e roxo as páginas do CCNH separadas por diferentes modularidades.

As comunidades do CCNH são divididas em três principais partes. Em roxo, encontram-se páginas gerais, relacionadas à pessoas, mural inicial, administração e notícias. Em verde, apenas páginas gerais. Em azul, páginas sobre a missão do CCNH, sobre pessoas e páginas externas.

Com relação ao CMCC, a comunidade em rosa representa principalmente as páginas relacionadas à diretoria, ensino e extensão, enquanto a comunidade preta se diferencia por páginas relacionadas ao perfil de docentes.



**Figura 3:** Grafo de hiperlinks dos centros da UFABC. Fonte: Figura gerada pelos autores do projeto no programa Gephi

No dia 25 de Junho de 2017 foram coletados os dados para o CMCC. O grafo gerado possui 869 páginas internas e 10 páginas externas relacionadas ao centro, chamadas de vértices, totalizando 879 vértices no grafo e 29493 arestas que representam as ligações entre essas páginas. Quanto ao número de arquivos, o centro possui 1056 arquivos no total, dos quais, 1044 são no formato PDF. Neste grafo, a página que possui o maior grau de saída, ou seja, que mais se refere a outras páginas internas do centro e externas é: [cmcc.ufabc.edu.br/index.php?option=com\\_sppagebuilder&view=page&id=148](http://cmcc.ufabc.edu.br/index.php?option=com_sppagebuilder&view=page&id=148). A página que possui maior grau de entrada, ou seja, mais citada e, portanto, mais central é: [http://cmcc.ufabc.edu.br/index.php?option=com\\_sppagebuilder&view=page&id=274](http://cmcc.ufabc.edu.br/index.php?option=com_sppagebuilder&view=page&id=274)

No dia 26 de Junho de 2017 foram coletados os dados para o CCNH. No grafo gerado foram identificadas 4511 páginas internas e 53 páginas externas à UFABC, totalizando 4564 vértices e 52777 arestas ligando estas páginas. Dos arquivos analisados, observamos que este centro possui 77332 arquivos, dos quais, 73737 correspondem a arquivos do formato PDF. Neste grafo, a página que possui o maior grau de saída, ou seja, que mais se refere a outras páginas internas e externas é: [ccnh.ufabc.edu.br/pessoas/docentes](http://ccnh.ufabc.edu.br/pessoas/docentes). A página que possui maior grau de entrada, ou seja, mais citada e, portanto, mais central é: <http://ccnh.ufabc.edu.br>

Por fim, no dia 25 de Junho de 2017 foram coletados os dados para o CECS. No grafo, foram identificadas 439 páginas internas e 19 páginas externas citadas pelas páginas das universidade, totalizando 458 páginas e 23993 arestas conectando essas páginas. Dos arquivos analisados, observamos que este centro possui 6349 arquivos no total, dos quais, 2277 correspondem a arquivos do formato PDF e 3574 do formato HTM. Neste grafo, a página que possui maior grau de saída é: [cecs.ufabc.edu.br/index.php/docentes/contatos.html](http://cecs.ufabc.edu.br/index.php/docentes/contatos.html). As páginas que possuem maior grau de entrada, ou seja, mais citadas e, portanto, mais centrais são: [cecs.ufabc.edu.br/dac/index.php/estagios.html](http://cecs.ufabc.edu.br/dac/index.php/estagios.html) e [cecs.ufabc.edu.br/index.php/tg.html](http://cecs.ufabc.edu.br/index.php/tg.html)

Neste projeto analisamos também os centros da UFABC separadamente e calculamos métricas globais e locais para cada um deles, bem como, analisamos os números de arquivos em cada página de cada centro. As métricas globais para o CCNH estão apresentadas na Tabela 3.

**Tabela 3:** Métricas globais do grafo de hiperlinks do CCNH.

	Grau			Mod.	Dens.	Agr.	Págs.
	Entrada	Saída	Total				
<b>Total</b>	43679	52778	96457	40989	0,003	0,531	4564
<b>Média</b>	9,682	11,699	21,382	9,086	—	—	—
<b>Desvio padrão</b>	82,350	7,543	83,913	3,614	—	—	—

Os resultados obtidos para os arquivos analisados em cada página deste centro estão apresentados na Tabela 4.

**Tabela 4:** Métricas globais referentes aos arquivos analisados do grafo de hiperlinks do CCNH.

Tipos de arquivos	Total	Média	Desvio Padrão
<b>jpg</b>	3389	0,751	0,432
<b>zip</b>	1	0	0,014
<b>ppt</b>	0	0	0
<b>htm</b>	40	0,008	0,126
<b>exe</b>	0	0	0
<b>txt</b>	0	0	0
<b>mp4</b>	0	0	0
<b>mp3</b>	3	0	0,044
<b>doc</b>	78	0,017	0,244
<b>pdf</b>	73737	16,346	2,721
<b>xls</b>	84	0,018	0,342

Na Tabela 5, apresentamos as métricas globais para o CECS. Na Tabela 6, apresentamos os resultados obtidos para os arquivos analisados em cada página deste centro.

**Tabela 5:** Métricas globais do grafo de hiperlinks do CECS.

	Grau						
	Entrada	Saída	Total	Mod.	Dens.	Agr.	Pág.
<b>Total</b>	22327	23989	46316	1070	0,115	0,701	458
<b>Média</b>	50,858	54,644	105,503	2,437	–	–	–
<b>Desvio padrão</b>	129,002	12,026	129,834	1,735	–	–	–

**Tabela 6:** Métricas globais referentes aos arquivos analisados do grafo de hiperlinks do CECS.

Tipos de arquivos	Total	Média	Desvio Padrão
<b>jpg</b>	10	0,022	0,287
<b>zip</b>	0	0	0
<b>ppt</b>	0	0	0
<b>htm</b>	3574	8,141	0,564
<b>exe</b>	0	0	15,955
<b>txt</b>	0	0	0
<b>mp4</b>	0	0	0,177
<b>mp3</b>	6	0,013	0,137
<b>doc</b>	63	0,143	0
<b>pdf</b>	2277	5,186	0,734
<b>xls</b>	419	0,954	0,208

Na Tabela 7, apresentamos as métricas globais para o CMCC. Na Tabela 8, apresentamos os resultados obtidos para os arquivos analisados em cada página deste centro.

**Tabela 7:** Métricas globais do grafo de hiperlinks do CMCC.

	Grau						
	Entrada	Saída	Total	Mod.	Dens.	Agr.	Págs.
<b>Total</b>	24986	29493	54479	605,595	0,038	0,690	879
<b>Média</b>	28,752	33,939	62,691	0,696	–	–	–
<b>Desvio padrão</b>	117,716	1,104	117,53	0,154	–	–	–

**Tabela 8:** Métricas globais referentes aos arquivos analisados do grafo de hiperlinks do CMCC.

Tipos de arquivos	Total	Média	Desvio Padrão
<b>jpg</b>	0	0	0,287
<b>zip</b>	0	0	0
<b>ppt</b>	0	0	0
<b>htm</b>	4	8,141	0,564
<b>exe</b>	0	0	15,955
<b>txt</b>	0	0	0
<b>mp4</b>	0	0	0,177
<b>mp3</b>	0	0	0,137
<b>doc</b>	8	0,143	0
<b>pdf</b>	1044	5,186	0,734
<b>xls</b>	0	0	0,208

Quanto ao número de páginas externas citadas pelas páginas do domínio da UFABC, temos como métricas globais os valores apresentados na tabela 9. O CCNH apresenta maior quantidade de links para páginas externas (ao todo 9085). Já o CECS apresenta a menor quantidade de links para páginas externas (ao todo 1662).

**Tabela 9:** Métricas referentes as páginas externas referenciadas nos centros CCNH, CECS e CMCC.

	Grau			Páginas
	Total	Média	Desvio Padrão	
<b>CCNH</b>	9085	171,415	761,279	53
<b>CECS</b>	1662	87,473	156,806	19
<b>CMCC</b>	4507	450,7	413,923	10

Observando as Tabelas 3, 4, 5, 6, 7 e 8 apresentadas, podemos evidenciar que as páginas do CCNH possuem mais arquivos do que os outros dois centros, principalmente na quantidade de PDFs e de arquivos de imagem, apesar de as páginas do CECS apresentarem um número maior de arquivos do formato HTM.

O grafo do CCNH é o que possui mais páginas e portanto um grau maior de entrada e saída. Porém, o grafo possui uma densidade de rede igual a 0,003 que pode ser considerada baixa, ou seja, as páginas do CCNH, apesar de serem muitas, são pouco conectadas entre si. A densidade do grafo do CECS é 0,115 e do CMCC 0,038. Apesar de também serem densidades de rede baixas, pode-se concluir que as páginas do CECS são mais conectadas quando comparadas com os outros dois centros e, portanto, podem ser consideradas as mais auto-suficientes e descriptivas nesse quesito. No Apêndice B é possível verificar os grafos das

páginas dos três centros separadamente.

Com relação ao coeficiente de agrupamento, o CECS é o centro com maior coeficiente igual a 0,701, o que significa que grande parte das páginas do centro estão conectadas entre si.

Por fim, analisando os números referentes às páginas externas na Tabela 9, o CCNH é o centro que possui maior número de páginas totais, o que significa que é o centro que mais se baseia em páginas externas ao domínio da UFABC para se estruturar na web. Em contrapartida, é possível afirmar que o CMCC é o centro mais auto-suficiente, ou seja, que menos depende de páginas externas no espaço web. Para o CCNH, a página externa com maior grau de entrada é [sites.google.com](http://sites.google.com). Para o CECS, [brasil.gov.br](http://brasil.gov.br) e [epwg.governoeletronico.gov.br](http://epwg.governoeletronico.gov.br). Para o CMCC, [lattes.cnpq.br](http://lattes.cnpq.br).

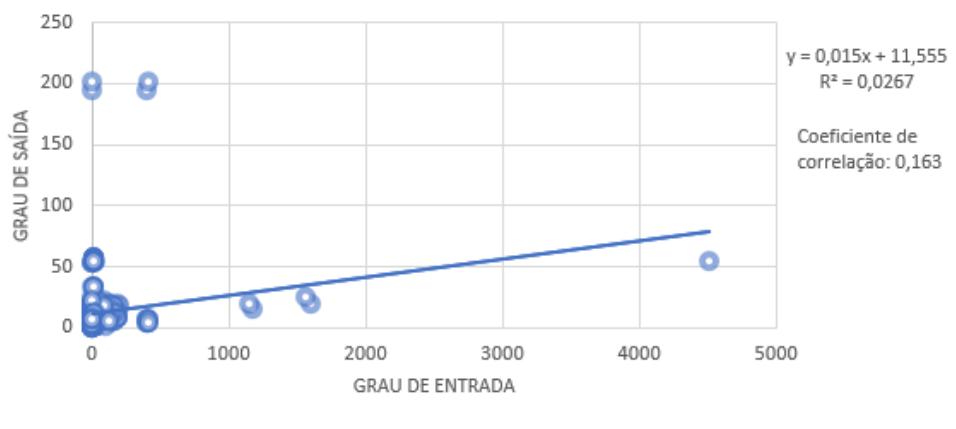
Na análise de métricas locais, observamos principalmente o grau de saída, grau de entrada e a modularidade dos grafos. Fizemos, então, uma comparação entre os graus dos três centros. Na Figura 4(a) representamos a relação entre os graus de entrada e de saída das páginas do CCNH, com destaque para a reta de regressão e o coeficiente de correlação entre as medidas.

Da mesma forma, na Figura 4(b) representamos a relação entre os graus de entrada e de saída das páginas do CECS, com destaque para a reta de regressão e o coeficiente de correlação entre as medidas. O mesmo foi feito para as páginas do CMCC na Figura 4(c)

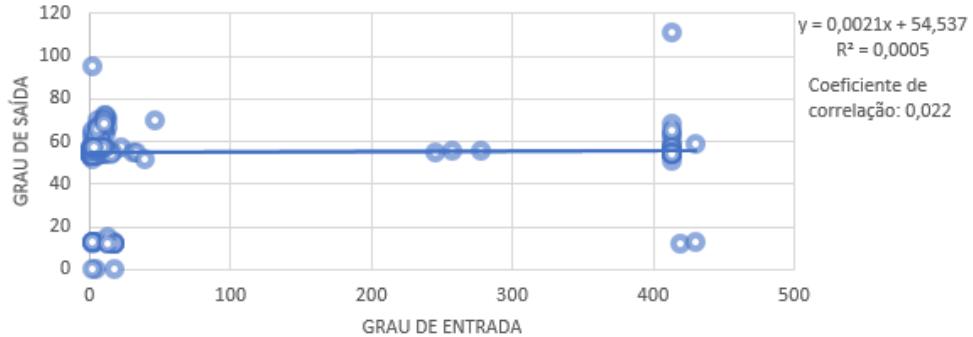
Da Figura 4 podemos observar que o coeficiente de correlação calculado pode ser considerado baixo, o que nos diz que a relação de dependência entre o grau de entrada e de saída das páginas de cada centro é pouco representativa.

Na Figura 4(a), é possível observar no gráfico que a grande maioria das páginas do CCNH possuem grau de entrada e de saída baixos, evidenciando que o centro tem páginas pouco conectadas e possui poucas páginas com grande visibilidade e centrais, ou seja, com alto grau de entrada. O grau de entrada baixo nessas páginas evidencia que elas tem pouca relevância para o espaço web da universidade.

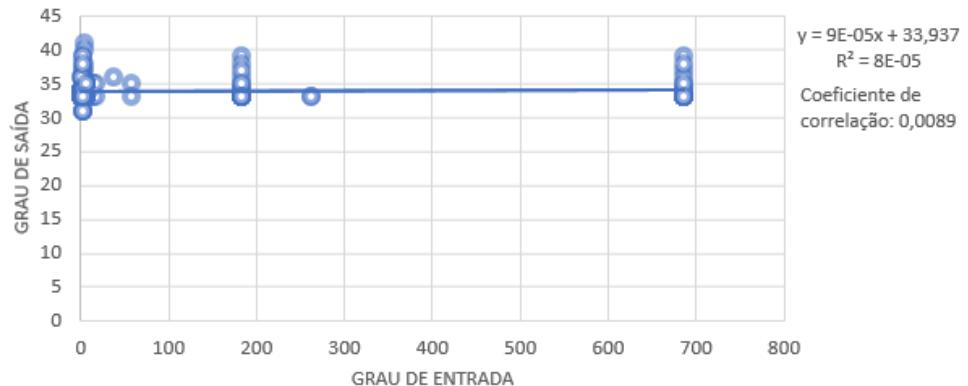
Com relação as Figuras 4(b) e 4(c), é possível observar nos gráficos que o CECS e o CMCC possuem muitas páginas com grau de entrada baixo e grau de saída alto, ou seja, são pouco citadas mas citam muitas páginas do centro ou externas tendo assim, um grau de importância baixo para a estruturação dos centros na web. Além disso, os centros possuem um número mediano de páginas com grau de entrada alto, ou seja, páginas centrais muito referenciadas por outras páginas.



(a) CCNH



(b) CECS



(c) CMCC

**Figura 4:** Gráfico de dispersão dos graus de entrada e de saída para os três centros da UFABC. Para cada figura representamos uma reta de regressão e seu coeficiente de correlação.

## 6 Conclusões

Para o desenvolvimento deste projeto de PDPD foram utilizadas diferentes ferramentas computacionais. Em particular usamos Python como linguagem de programação para a criação do grafo, LibreOffice Cal como plataforma para análise estatística, e o Gephi como ferramenta de visualização de grafos.

Com o objetivo de explorar a dinâmica complexa presente no espaço web da UFABC, é possível evidenciar que o centro de maior tamanho é o CCNH, com 4564 páginas totais e 77332 arquivos, sendo a maioria no formato PDF. Por outro lado, este centro também possui muitas páginas com grau de entrada e de saída baixos, ou seja, que são de pouca relevância para a estruturação do centro na web. Quanto a densidade do grafo, o CCNH é o centro de menor densidade de rede, ou seja, possui páginas pouco conectadas.

Ao estudar a conectividade dos centros, pode-se observar que o CECS é o que possui mais páginas conectadas, ou seja, pode ser considerado mais auto-suficiente e descriptivo. Isso pode ser evidenciado pelo fato de o CECS ser o centro de maior densidade de rede e maior coeficiente de agrupamento. Além disso, o CECS e o CMCC são os centros que possuem mais páginas com grau de entrada altos, ou seja, possuem páginas centrais relevantes para o espaço web.

Por fim, ao considerar o estudo de páginas externas do centro, observa-se que o centro que mais cita páginas externas é o CCNH, enquanto o que menos referencia páginas externas é o CMCC.

Este projeto que estamos concluindo permitiu, além de seguirmos o método de pesquisa científica (que envolve aspectos desde a definição do problema até a sua solução e reflexão), desenvolvemos capacidades técnicas que serão muito úteis para futuros empreendimentos na vida acadêmica ou profissional.



## A Algoritmo para criação do grafo de hiperlinks

```
1 import sys
2 import shutil
3 import fileinput
4 import unicodedata
5
6 def indicePagina(url, vertices, host):
7     url = url.strip()
8     for i in range(0, len(vertices)):
9         if url==vertices[i][0]:
10             return i
11     if (not host in url):
12         url = url.replace('http://','')
13         url = url.replace('https://','')
14         url = url.replace('www. ','')
15         url = url.split('/')[-1]
16         for i in range(0, len(vertices)):
17             if vertices[i][1]=='externo' and url==vertices[i][0]:
18                 return i
19
20     elif url.endswith('.pdf'):
21         return "pdf"
22     elif url.endswith(".htm") or url.endswith(".html") or url.endswith(".oth") or
url.endswith(".php") :
23         return "htm"
24     elif url.endswith(".doc") or url.endswith(".docx") or url.endswith(".rtf") or
url.endswith(".odt") :
25         return "doc"
26     elif url.endswith(".xls") or url.endswith(".xlsx") or url.endswith(".ods") or
url.endswith(".odp") :
27         return "xls"
28     elif url.endswith(".ppt") or url.endswith(".pptx") or url.endswith(".odp") :
29         return "ppt"
30     elif url.endswith(".zip") or url.endswith(".rar") or url.endswith(".tgr") or
url.endswith(".tar.gz") or url.endswith(".7z") or url.endswith(".tar") or url.
endswith(".xz") :
31         return "zip"
32     elif url.endswith(".txt") :
33         return "txt"
34     elif url.endswith(".jpg") or url.endswith(".jpeg") or url.endswith(".bmp") or
url.endswith(".png") or url.endswith(".gif") or url.endswith(".tif") :
35         return "jpg"
36     elif url.endswith(".mp3") or url.endswith(".wma") or url.endswith(".aac") or
url.endswith(".wav") or url.endswith(".ac3") :
37         return "mp3"
38     elif url.endswith(".mp4") or url.endswith(".avi") or url.endswith(".mpeg") or
url.endswith(".mov") or url.endswith(".rmvb") :
39         return "mp4"
40     elif url.endswith(".exe") or url.endswith(".bin") or url.endswith(".sh") :
41         return "exe"
42     else:
43         return -1
44
45
```

```

46 # PROGRAMA PRINCIPAL
47 if __name__ == "__main__":
48     print("Iniciando leitura do arquivo.")
49     arquivoEntrada = "cecs.txt"
50     host = "ufabc.edu.br"
51
52     # FASE 1: Criacao de lista de vertices
53     count = 0
54     vertices = list([])
55     arestas = list([])
56     listaDeExtensoes = ['pdf', 'htm', 'doc', 'xls', 'ppt', 'zip', 'txt', 'jpg',
57     'mp3', 'mp4', 'exe']
58
59     for line in fileinput.input(arquivoEntrada):
60         line = line.strip('\n')
61         line = line.replace("//", "/")
62         if ("main_html" in line):
63             line = line[10:]
64             count = count+1
65             contadorDeArquivos = dict([])
66             for ext in listaDeExtensoes:
67                 contadorDeArquivos[ext] = 0
68             vertices.append((line, 'interno', contadorDeArquivos))
69         else:
70             if line.startswith("\t") and (not host in line):
71                 line = line.strip()
72                 line = line.replace("http://", "")
73                 line = line.replace("https://", "")
74                 line = line.replace("www.", "")
75                 line = line.replace("wwws.", "")
76                 line = line.replace("www1.", "")
77                 line = line.split("/")[0]
78                 if (not (line, 'externo', []) in vertices):
79                     vertices.append((line, 'externo', []))
80
81     # FASE 2: Criacao de lista de arestas
82     for line in fileinput.input(arquivoEntrada):
83         line = line.strip('\n')
84         line = line.replace("//", "/")
85         if ("main_html" in line):
86             line = line[10:]
87             indiceOrigem = indicePagina(line, vertices, host)
88         elif line.startswith("\t"):
89             line = line.replace("\t", "")
90             indiceDestino = indicePagina(line, vertices, host)
91             if indiceDestino in listaDeExtensoes:
92                 vertices[indiceOrigem][2][indiceDestino] += 1
93             elif not indiceDestino== -1:
94
95     # FASE 3: Criacao do grafo
96     arquivoSaida = arquivoEntrada + ".gdf"
97
98     s = "nodedef>name VARCHAR, label VARCHAR, tipo VARCHAR, numeropdfs DOUBLE,
99     numerohtm DOUBLE, numerodoc DOUBLE, numeroxls DOUBLE, numeroppt DOUBLE,
100    numerozip DOUBLE, numerotxt DOUBLE, numerojpg DOUBLE, numeromp3 DOUBLE,
101    numeromp4 DOUBLE, numeroexe DOUBLE"

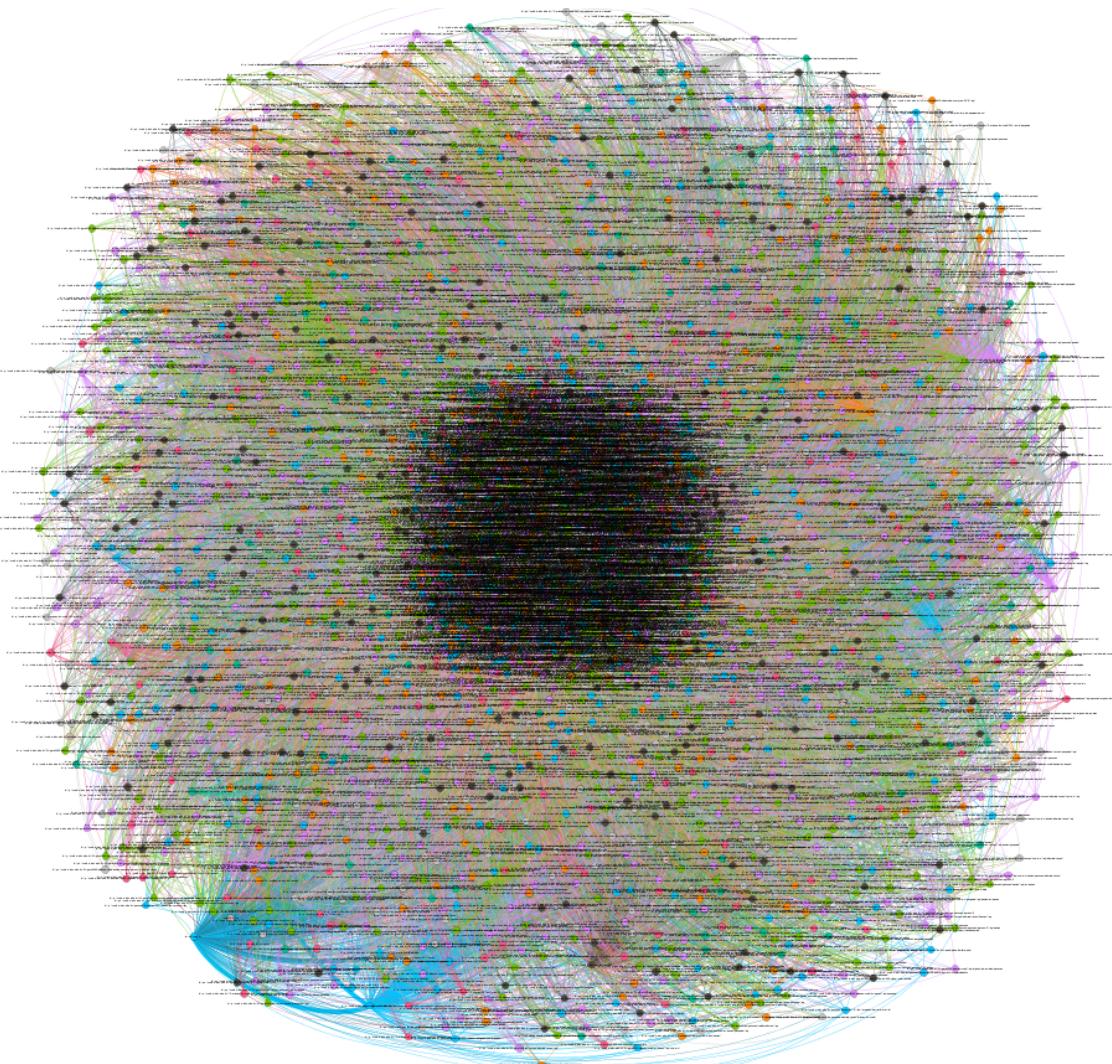
```

```
98
99     for i in range(0,len(vertices)):
100         s += "\n" + str(i) + "," + vertices[i][0]+ "," + vertices[i][1]
101         for ext in listaDeExtensoes:
102             if vertices[i][1]== 'interno':
103                 s += "," + str(vertices[i][2][ext])
104             else:
105                 s += ",0"
106         s += "\nedgedef>node1 VARCHAR, node2 VARCHAR, directed BOOLEAN"
107         for i in range(0,len(arestas)):
108             s += "\n" + str(arestas[i][0]) + "," + str(arestas[i][1]) + ", true"
109         output = open(arquivoSaida, 'w')
110         output.write(s)
111         output.close
```

## B Grafos de hiperlinks do *website* da UFABC

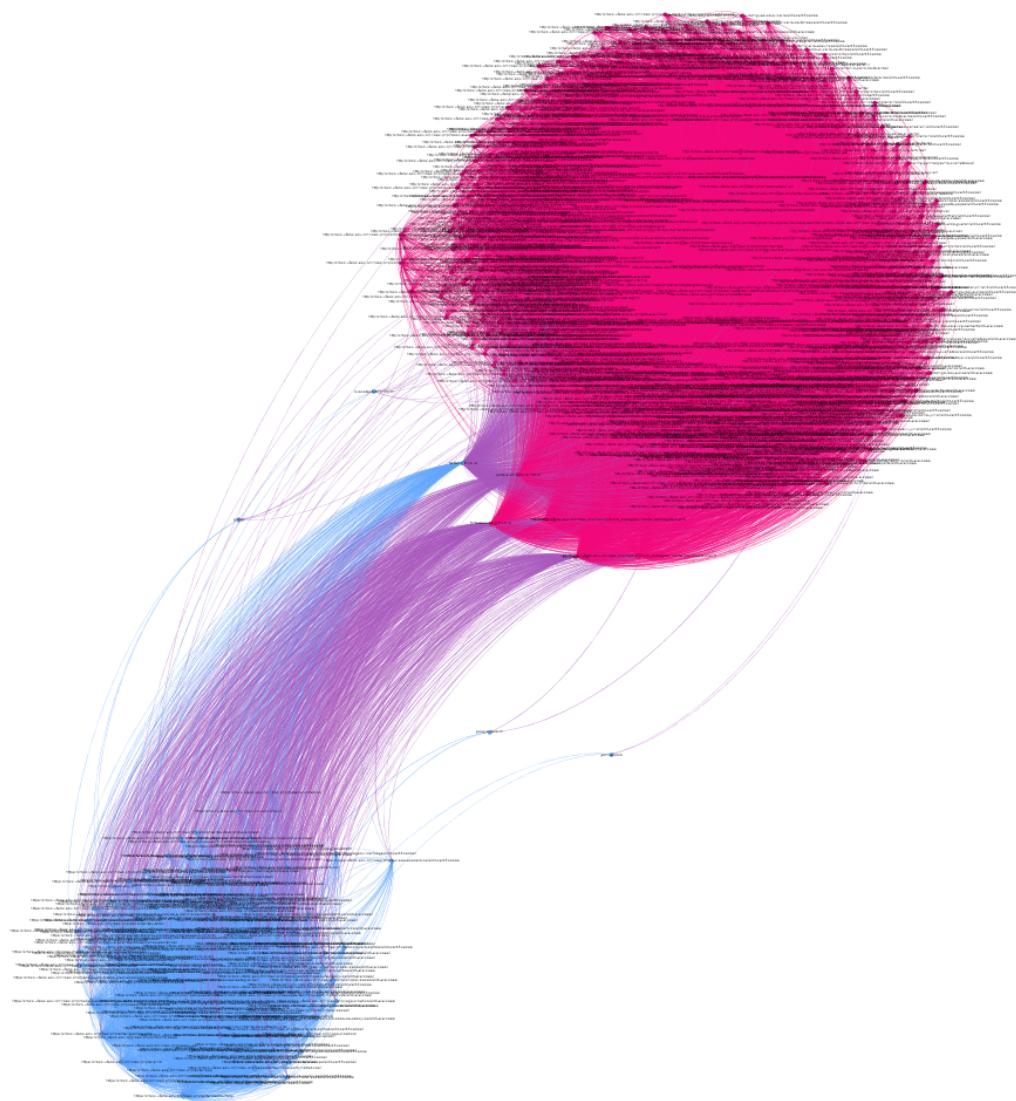
Nas Figuras 5, 6, 7 apresentamos o grafo do CCNH, CMCC e CECS, respectivamente.

Na Figura 5 não é possível observar comunidades bem definidas. Isso se da pelo fato de as páginas do site estarem bem conectadas a ponto de não serem capazes de formarem comunidades específicas como nos outros grafos.



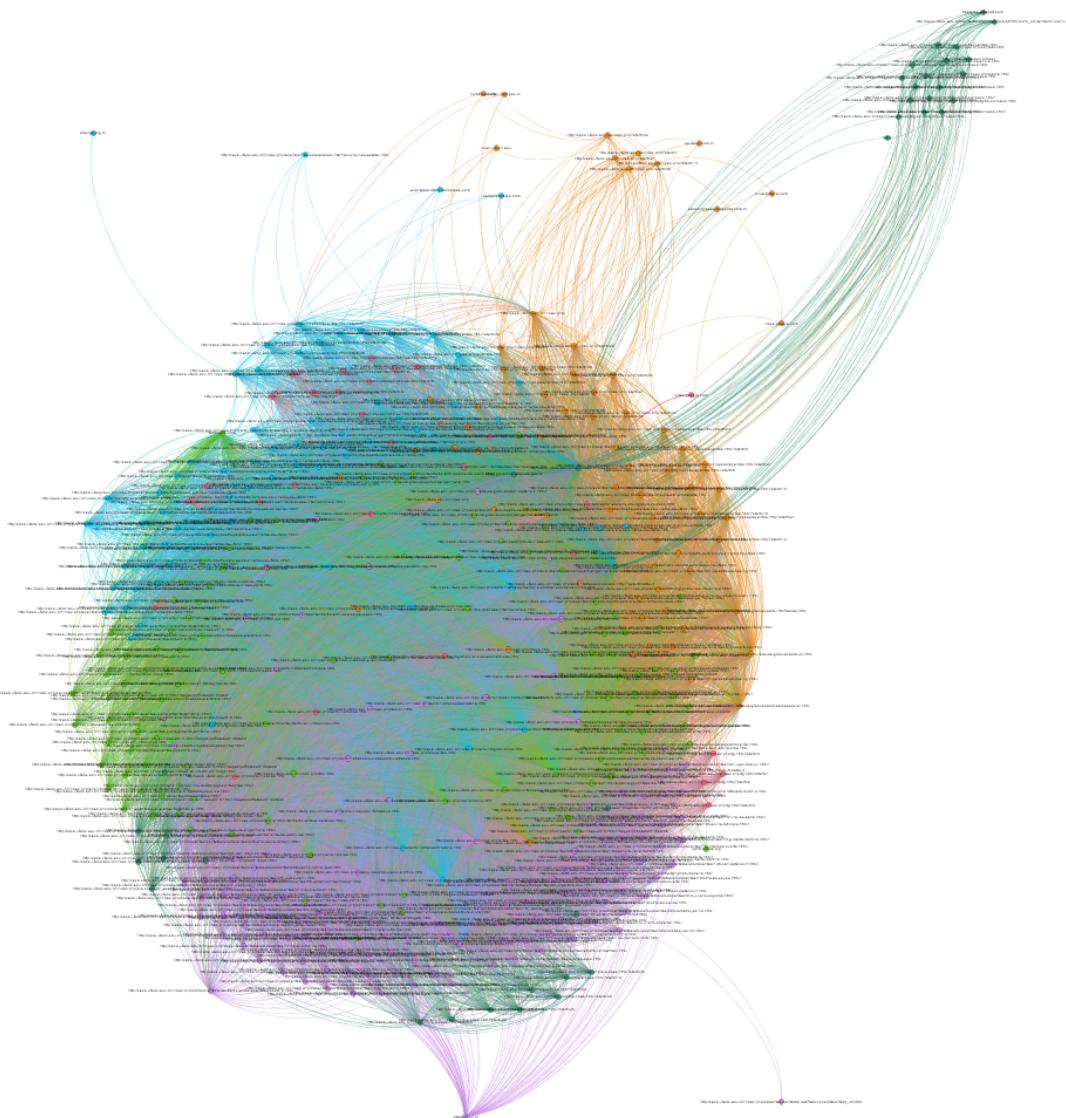
**Figura 5:** Grafo de hiperlinks da página do CCNH. As cores foram definidas de acordo com a modularidade. Fonte: Figura gerada pelos autores do projeto no programa Gephi.

Na Figura 6 é possível observar duas comunidades bem divididas no grafo obtido do CMCC. Essa divisão se dá pelo fato de, em rosa, estarem as páginas relacionadas principalmente a perfis de professores do centro, certificados e eleições, enquanto em azul, temos principalmente páginas centrais do CMCC, como por exemplo as páginas da divisão acadêmica, diretoria, estagiários, técnicos administrativos, conselho. Os nós centrais que ligam essas duas comunidades são páginas externas.



**Figura 6:** Grafo de hiperlinks da página do CMCC. As cores foram definidas de acordo com a modularidade. Fonte: Figura gerada pelos autores do projeto no programa Gephi.

Na Figura 7, apesar de não existirem comunidades bem definidas como na anterior, é possível observar dois destaques no grafo obtido para o CECS. No canto superior direito, em verde, é possível identificar uma comunidade que se destaca por ser só de páginas do conselho do centro (ConCECS) e da divisão acadêmica, além de uma página externa sobre *templates*. Na parte inferior, em roxo, é possível observar uma página muito referenciada, que é a página [lattes.cnpq.br](#). As outras páginas destacadas fora do centro do grafo são páginas externas.



**Figura 7:** Grafo de hiperlinks da página do CECS. As cores foram definidas de acordo com a modularidade. Fonte: Figura gerada pelos autores do projeto no programa Gephi.

### Referências

- Bastian, M., Heymann, S., Jacomy, M., et al. (2009). Gephi: an open source software for exploring and manipulating networks. *Icwsom*, 8:361–362.
- da Silva, I. C. O. (2011). Aplicação de indicadores webométricos nos programas de pós-graduação das engenharias recomendados pela capes. Master's thesis, Universidade Federal do Rio Grande do Norte.
- Figueiredo Filho, D. B. and Silva Junior, J. A. (2010). Desvendando os mistérios do coeficiente de correlação de pearson (r). *Revista Política Hoje-ISSN: 0104-7094*, 18(1):115–146.
- Gouveia, F. C. (2012). Novos caminhos e alternativas para a webometria. *Em Questão*, 18(3):249–261.
- Jeyashree, S. and Ravichandran, R. (2013). Perspectives of webometric tools for web impact assessment studies: A review. *International Journal of Library Science*, 2(2):43–48.
- Orduña-Malea, E. and Agullo, I. F. (2015). *Cibermetría. Midiendo el espacio red*, volume 1. Editorial UOC.
- Thelwall, M. (2009). *Introduction to Webometrics: Quantitative Web Research for the Social Sciences*, volume 4. Morgan & Claypool Publishers.
- Thelwall, M. (2012). A history of webometrics. *Bulletin of the Association for Information Science and Technology*, 38(6):18–23.
- Vanti, N. A. P. (2002). Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento. *Ciência da informação*, 31(2):152–162.
- Vanti, N. A. P. (2007). Aplicação de indicadores web aos sites acadêmicos latino-americanos em ciências sociais. *Brazilian Journal of Information Science*, 1(2):22–46.