*Ministério da Educação*
**Universidade Tecnológica Federal do Paraná**
*Pró-Reitoria de Pesquisa e Pós-Graduação*

**Relatório Final de Atividades**

# RECONHECIMENTO DE PADRÕES EM REDES DE COAUTORIA UTILIZANDO REDES COMPLEXAS

**vinculado ao projeto**

**Reconhecimento de padrões em sequências genômicas: um estudo de caso utilizando redes complexas**

**Matheus Montanini Breve**

**Bolsista UTFPR**

**Engenharia Elétrica**

**Data de ingresso no programa: 09/2014**

**Prof. Dr. Fabrício Martins Lopes**

Área do Conhecimento: 1.00.00.00-3 Ciências Exatas e da Terra

*CAMPUS CORNÉLIO PROCÓPIO*, 2015

**MATHEUS MONTANINI BREVE**
**FABRÍCIO MARTINS LOPES**

# RECONHECIMENTO DE PADRÕES EM REDES DE COAUTORIA UTILIZANDO REDES COMPLEXAS

Relatório de pesquisa do Programa de Iniciação Científica da Universidade Tecnológica Federal do Paraná sob orientação do Prof. Dr. Fabrício Martins Lopes.

CORNÉLIO PROCÓPIO, 2015

**SUMÁRIO**

## INTRODUÇÃO

A teoria dos grafos, base para a teoria de redes, originou-se em 1736 a partir de Leonhard Euler com a sua solução para o problema das sete pontes de Königsberg, antes uma cidade do reino da Prússia que foi renomeada para Kaliningrado pela União Soviética. Esta cidade é cortada pelo rio Pregolya, que formava uma composição geográfica com quatro porções de terra ligadas por sete pontes. Naquela época, os habitantes de Königsberg se perguntavam se era possível cruzar as sete pontes sem cruzar a mesma ponte duas vezes. Euler propôs uma solução envolvendo o que agora denominamos teoria dos grafos, representando cada porção de terra por um nó e cada ponte entre as porções de terra por uma conexão entre os respectivos nós. Matematicamente, Euler provou que este caminho não existia, pois havia um número ímpar de conexões entre as porções de terra [1].

Desde então, a teoria dos grafos desenvolveu-se significantemente com importantes contribuições de vários pesquisadores, principalmente matemáticos, que tentavam criar soluções para vários problemas modelando-os a partir de grafos. Uma das consequências deste desenvolvimento foi o surgimento da teoria de redes.

Após o surgimento e consequente aprofundamento da teoria das redes, estas começaram a ser empregadas em diversos problemas, como em biologia, na identificação de proteínas ou genes importantes para um determinado processo biológico, ou até mesmo em segurança e criptografia na *internet*, verificando a robustez e tolerância a ataques nas redes da *internet*. As redes são omnipresentes e possuem enorme valor em inúmeras áreas do conhecimento. Em relação as redes sociais, este trabalho se refere as redes de coautoria entre pesquisadores, área que vem ganhando destaque devido a crescente quantidade de informação presente na rede em plataformas como *Google Scholar*, *ResearchGate* e a Plataforma *Lattes* no Brasil.

Alguns trabalhos já foram publicados abordando essas redes de coautoria, com enfoque no Brasil, como por exemplo o artigo [2], no qual a rede formada pelos cursos de pós-graduação na área de Ciência da Computação foi mapeada e analisada em relação a produtividade acadêmica e o artigo [3], onde as redes formada por grupos de pesquisadores contidos na Plataforma *Lattes* foram analisadas com o intuito de caracterizá-las topologicamente.

Dada a importância das redes nas mais diversas áreas e a possibilidade de utilizá-las para analisar as redes de autoria, foram executados diversos procedimentos realizados entre Setembro de 2014 e Julho de 2015 para a obter dados sobre os pesquisadores de cada Programa de Pós-Graduação e Extensão (PPGE) de forma a gerar redes de coautoria, permitindo assim a posterior análise das redes entre os pesquisadores e as instituições dos PPGEs brasileiros utilizando métricas de redes complexas. De posse das redes, realizou-se o cálculo das métricas típicas de redes complexas, que foram calculadas através de programas específicos como o R [4] e o *Gephi* [5]. Com estas informações, o objetivo futuro do projeto é verificar padrões que determinam as características dos programas ou das áreas ao longo dos triênios de avaliação, como o conceito designado pelo órgão maior, neste caso, a CAPES, ou a produção acadêmica, por exemplo. Este relatório contém a descrição de todos os procedimentos realizados durante este período.

## METODOLOGIA

Esta seção contém detalhamentos sobre os procedimentos tomados para atingir os objetivos definidos na Seção \ref{sec:objetivos}. Os procedimentos serão divididos em duas partes.

A primeira parte comprime os processos realizados para a extração e armazenamento de dados para formar a base de dados contendo os pesquisadores, seus respectivos PPGEs e suas respectivas situações entre 2004 e 2012. A Figura 1 contém um fluxograma descrevendo simplificadamente os processos componentes da primeira parte.

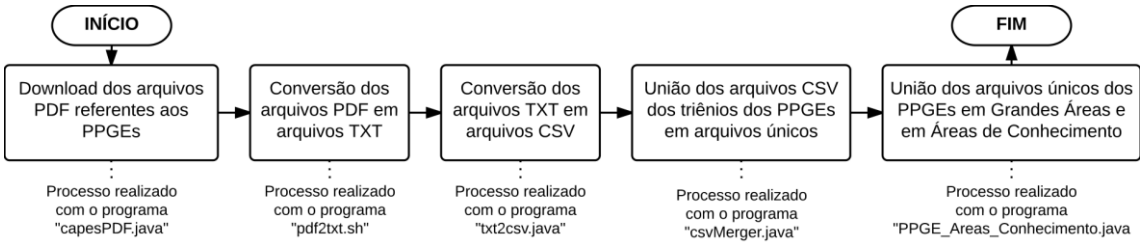


Figura 1. Processos executados na primeira parte

Para realizar o *download* dos arquivos PDF de avaliação referentes a cada PPGE e a cada triênio a partir do servidor da CAPES, foi utilizada uma lista de URLs desenvolvida por Jesús Pascual Mena-Chalco. Esta lista de URLs continha, no total, 23248 endereços, representando, assim, 23248 arquivos em formato PDF. Os arquivos em formato PDF de avaliação dos PPGEs apresentam a forma da Figura 2 quando extraídos do portal da CAPES.

Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior

Sistema de Avaliação

Indicadores de Corpo Docente Atuação

40005011008P0- CIÊNCIAS / UEPG - 2012

| Nome Docente (1) | Ano Categoria (2) | Afastado | Disciplinas (3) | | Carga Horária | | Participação em projeto pesquisa (4) | | Orientação (5) | | | | | | | Formação no Programa | Participação em Banca (6) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Graduação | | | Pós Graduação | | | Concluídas | | |
| | | | Grad. | Pós Grad. | Grad. | Pós Grad. | Equipe | Resp. | Inic Cientifica | Tutoria | Monografia | Mes | Dou | Prof. | | | |
| Alcione Roberto Jurelo | 2010 P 2011 P 2012 P | | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 3 | 4 | 0 | 0 | Não | 0 |
| André Mauricio Brinatti | 2012 P | | 3 | 0 | 204 | 0 | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | Não | 2 |
| Antonio Marcos Batista | 2010 P 2011 P 2012 P | | 0 | 3 | 0 | 240,00 | 2 | 3 | 0 | 0 | 0 | 2 | 4 | 0 | 1 | Não | 0 |
| Carlos Eugênio Foerster | 2010 P 2011 P 2012 P | | 2 | 1 | 340 | 30,00 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Não | 0 |
| Fabio Augusto Meira Cássaro | 2011 P 2012 P | Estágio pós-doutoral | 0 | 0 | 0 | 0 | 4 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | Não | 1 |
| Francisco Carlos Serbena | 2010 P 2011 P 2012 P | Outro Motivo | 1 | 1 | 136 | 90,00 | 3 | 2 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | Não | 0 |
| Gelson Biscaia de Souza | 2012 C | | 2 | 0 | 408 | 0 | 0 | 1 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | Não | 0 |
| Jose Danilo Szezech Júnior | 2012 P | | 4 | 0 | 442 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | Não | 0 |
| Luiz Fernando Pires | 2010 P 2011 P 2012 P | | 1 | 1 | 204 | 45,00 | 2 | 3 | 1 | 0 | 0 | 2 | 1 | 0 | 0 | Não | 0 |
| Pedro Rodrigues Junior | 2010 P 2011 P 2012 P | | 2 | 0 | 272 | 0 | 2 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | Não | 1 |
| Sandro Ely de Souza Pinto | 2010 P 2011 P 2012 C | | 0 | 1 | 0 | 45,00 | 1 | 1 | 0 | 0 | 0 | 4 | 4 | 0 | 1 | Não | 1 |
| Sérgio da Costa Saab | 2010 P 2011 P 2012 P | | 3 | 0 | 272 | 0 | 0 | 1 | 2 | 0 | 0 | 2 | 2 | 0 | 1 | Não | 2 |
| Sergio Leonardo Gómez | 2010 P 2011 P 2012 P | | 1 | 3 | 408 | 150,00 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | Não | 0 |

Chronos 1 de 2 29/05/13 às 15:45

Figura 2. Formato dos arquivos PDF de avaliação do porta da CAPES

Por meio de um programa, denominado ``CapesPDF'' escrito em linguagem Java e com o auxílio da biblioteca *Apache Commons IO*, todos os endereços foram corretamente acessados e o seus conteúdos salvos automaticamente, exceto em casos como ``*Connection Timeout*'' ou arquivos inexistentes o corrompidos no servidor. Para estes arquivos corrompidos ou inexistentes no servidor, comprovou-se que se referiam a programas de pós-graduação e extensão desativados. O fluxograma que representa a ordem de ações tomadas pelo programa ``capesPDF'' encontra-se na Figura 3.

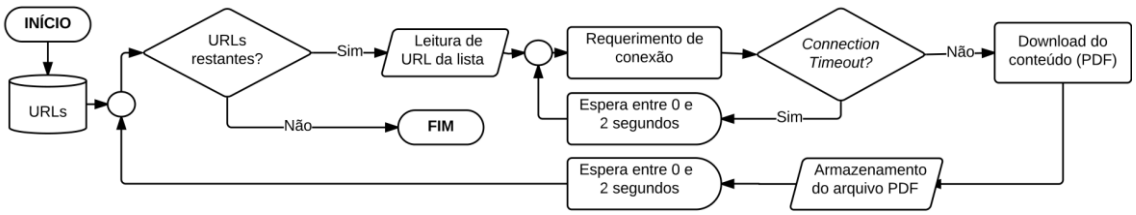


Figura 3. Processos executados para *download* dos arquivos PDF de avaliação

De forma a obter uma forma viável para a extração dos dados, optou-se por converter os arquivos em formato PDF das avaliações dos PPGEs para arquivos de texto, permitindo a atuação de um *parser* para extração das informações importantes. Para atingir tal objetivo, um *shell script* denominado ``pdf2txt.sh'' foi desenvolvido por Jesús Pascual Mena-Chalco para a conversão dos arquivos PDF em arquivos de texto.
Um exemplo de arquivo texto gerado a partir de um arquivo PDF encontra-se na Figura 4. Na figura é possível observar a existência do nome dos posquisadores e as suas

respectivas posições nos PPGEs de atuação nos anos do triênio em questão. As letras P, C e V utilizadas para denotar a participação dos pesquisadores nos PPGEs em cada ano significam, respectivamente: permanente, colaborador e visitante.

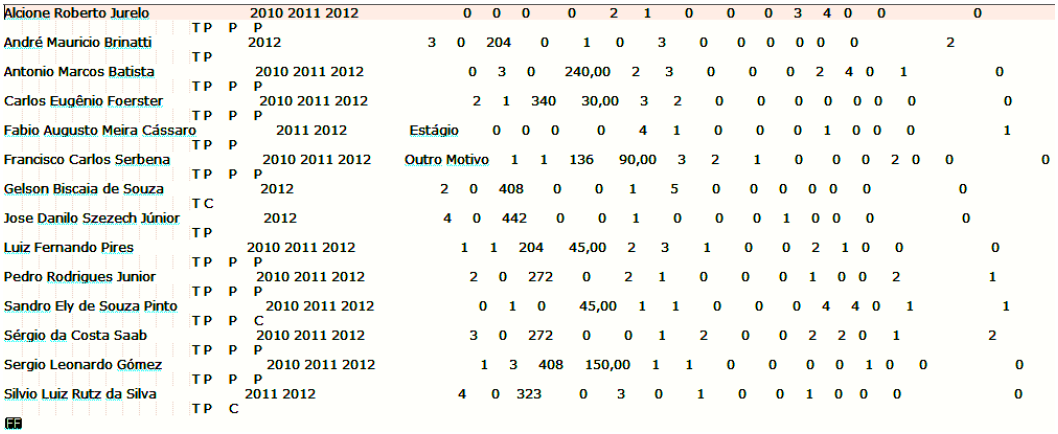


Figura 4. Formato dos dados em formato de texto (TXT)

Com o intuito de se adquirir um formato de dados apropriado para análise, criou-se um programa, denominado ``txt2csv'' em linguagem Java para a extração e geração automática de dados organizados em um arquivo CSV contendo o nome de cada pesquisador e a sua situação no PPGE em questão durante os anos da avaliação trienal. Um fluxograma contendo uma representação dos processos executados pelo programa encontra-se na Figura 5.

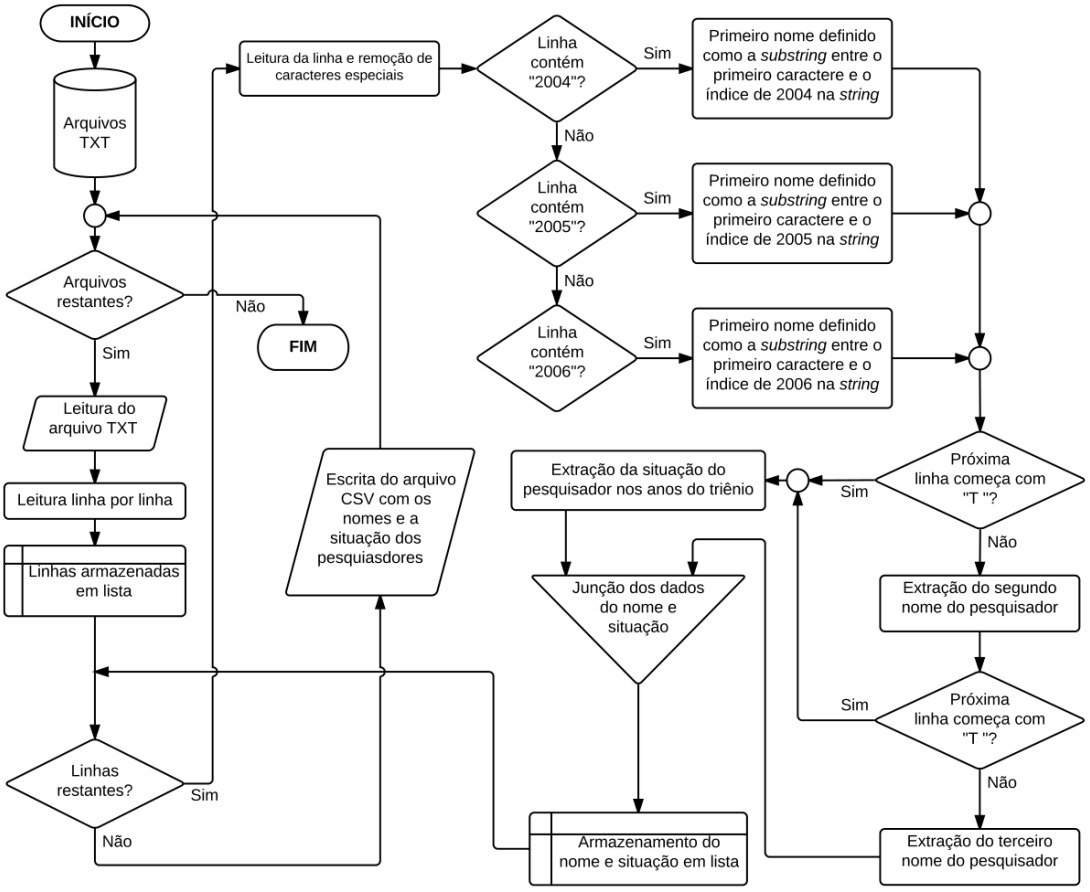


Figura 5. Processos pertinentes ao programa “txt2csv”

Após a execução do programa, todos os arquivos de texto foram convertidos para arquivos CSV, gerando arquivos para cada PPGE e para cada triênio avaliado do total de três triênios. Para condensar a informação dos pesquisadores dos PPGEs nos três triênios de forma a permitir a análise intra-programas das redes de coautoria, criou-se um programa em Java, denominado ``CsvMerger'', responsável por unir os arquivos CSV referentes aos três triênios, organizando assim os pesquisadores presentes no PPGE entre 2004 e 2012 e as suas respectivas situações em cada um dos anos da avaliação.

Além de condensar as informações referentes aos PPGEs, foi desenvolvido também um *script* para condensar os dados dos pesquisadores das áreas e grandes áreas do conhecimento, conforme definidas pela CAPES. O resultado deste programa, denominado ``AreasCsvMerger'', é a geração automática de arquivos CSV contendo as informações de todos os pesquisadores dos PPGEs de uma certa área de conhecimento, juntamente com a informação da sua situação nos anos de avaliação dos triênios. Um fluxograma contendo uma representação dos processos relativos ao programa "AreasCsvMerger" encontra na Figura 6.

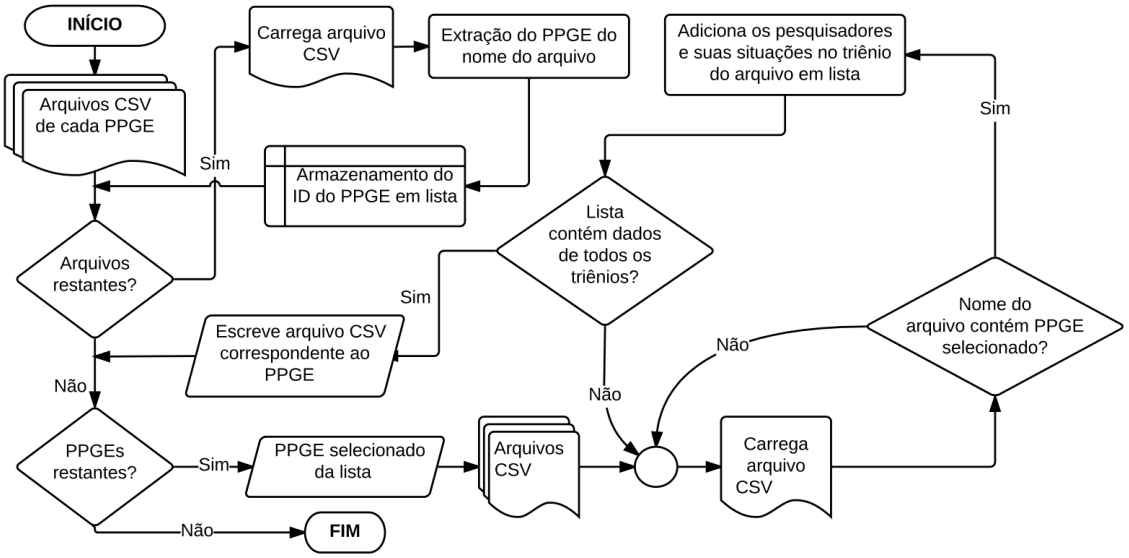


Figura 6. Programa "AreasCsvMerger"

Para cada arquivo gerado a partir dos procedimentos executados mencionados anteriormente, utilizou-se um programa para a obtenção dos números de identificação dos currículos Lattes dos pesquisadores listados em cada arquivo. Este programa, denominado ``busca-pessoas-BD-interno'', foi desenvolvido em linguagem de programação *Python* por Jesús Pascual Mena-Chalco.

O programa utiliza uma base de dados contendo os 3.820.292 pesquisadores registrados na plataforma Lattes até novembro de 2014 e os seus respectivos *Lattes ID*, em formato de 10 caracteres. O algoritmo desenvolvido e executado com o programa consiste na busca aproximada com base na similaridade entre dois nomes, mais precisamente, com base na distância de Levenshtein. Após sua execução em um arquivo de entrada contendo nomes de pesquisadores, o programa gera um arquivo contendo os nomes dos mesmos ao lado dos nomes mais similares encontrados na base de dados, juntamente com o *Lattes ID* do pesquisador identificado.

## RESULTADOS E DISCUSSÕES

Os arquivos contendo os dados dos pesquisadores foram divididos em PPGEs, Áreas do Conhecimento e Grandes Áreas. O objetivo de tal divisão é permitir a análise das redes de coautoria geradas a partir de cada PPGE (denominada intra-programas), entre os PPGEs (denominada entre-programas), nas áreas e grandes áreas do conhecimento. Para a geração dos arquivos de rede a partir dos dados dos esquisadores, utilizou-se o *software ScriptLattes*, desenvolvido por Jesús Mena-Chalco. O programa foi responsável pela geração dos arquivos de rede GML (Graph Markup Language) e GDF, formatos utilizados na biblioteca igraph na linguagem de programação R e no software Gephi, respectivamente.

Para cada categoria, isto é, PPGE, Área e Grande Área, os arquivos das redes são divididos em períodos, denominados ``p00", que representa o período de 2004 a 2012 que contém os três triênios da análise (rede global), ``p01" que representa o triênio 2004-2006, ``p02" o triênio 2007-2009 e ``p03" o triênio 2010-2012. Utilizando o software de visualização de redes Gephi, pode-se obter uma visão da topologia da rede além de, também, calcular algumas métricas de redes complexas. Um exemplo de visualização encontra-se na Figura 7, a qual contém a rede de coautoria entre pesquisadores da área de Engenharia no período global de avaliação (p00).

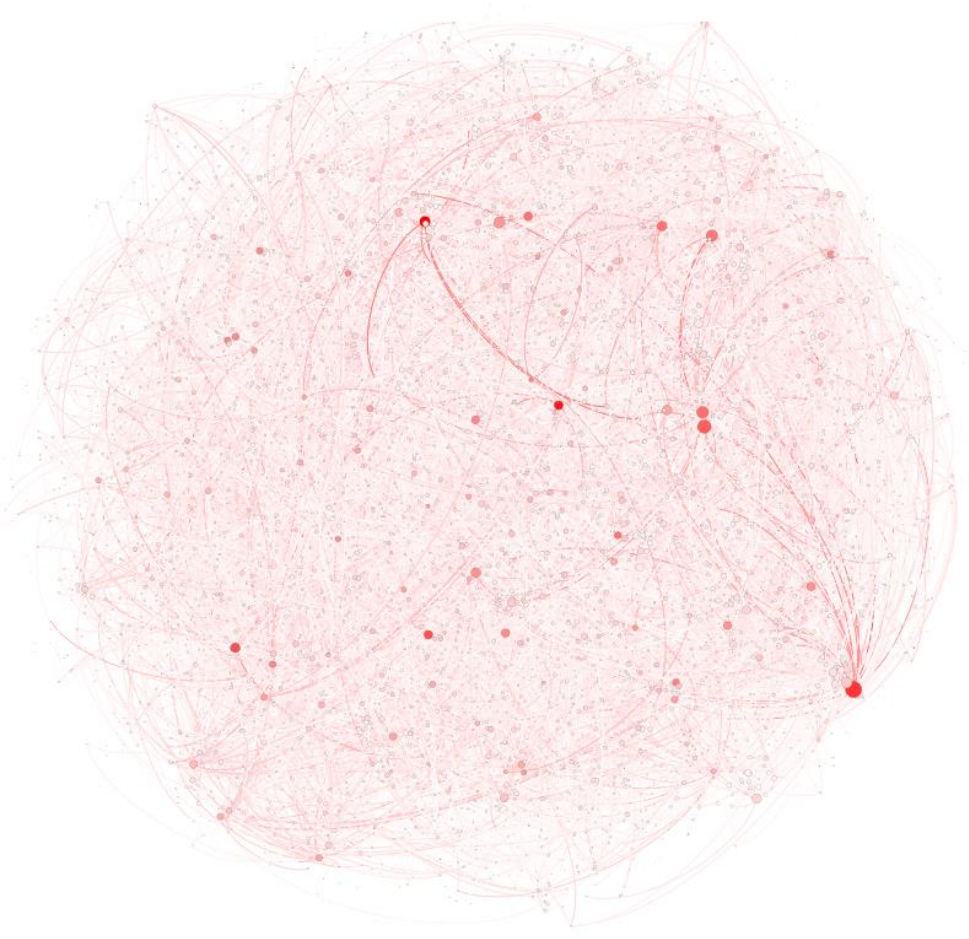

Figura 7. Rede de coautoria da área de Engenharia Elétrica de 2004 a 2012

Na figura anterior, os nós são coloridos de acordo com o número de conexões com outros nós em uma escala gradiente de vermelho, com o vermelho mais saturado sendo os nós com maiores números de conexões.

De forma a obter as métricas de rede, no total, 25 diferentes métricas, para as redes de cada PPGE, Área e Grande Área do conhecimento, utilizou-se a biblioteca *igraph* juntamente com a linguagem de programação R e a IDE *RStudio.* Algumas métricas geradas foram, por exemplo, o grau médio da rede, a transitividade, caminho médio e diâmetro da rede.

Após o cálculo das métricas, estas são salvas juntamente com a informação da área ou do PPGE em questão, como os conceitos CAPES nos anos de avaliação, a região e a situação jurídica do PPGE e outras informações além do triênio de referência.

Para as métricas referentes as Grandes Áreas e as Áreas do Conhecimento, não foi necessária a inclusão de informações extras no arquivo e, portanto, apenas adicionou-se o triênio de referência como uma coluna nestes casos. Foram gerados arquivos contendo as métricas para cada triênio e também para o período global de avaliação. Com o intuito de obter um arquivo único com as métricas de todos os triênios para cada categoria, isto é, PPGEs, Áreas e Grandes Áreas, utilizou-se o *software LibreOffice* e realizou-se a união dos arquivos manualmente, adicionando uma coluna contendo a informação sobre o triênio ao qual as métricas se referem, de forma a permitir a análise com base na evolução da área ou PPGE durante os triênios de avaliação.

## CONCLUSÕES

Os objetivos definidos inicialmente foram atingidos, obtendo como resultado a integração dos dados correspondentes aos pesquisadores analisados e consequentemente as redes de coautoria entre os mesmos, estes pertencentes a um mesmo programa de pós-graduação, a uma mesma área e a mesma grande área do conhecimento.

No total, foram gerados 34938 arquivos contendo os dados de cada rede, arquivos tanto em formato para uso no *software* R e para uso no *software Gephi*, divididos em períodos. Tais arquivos permitiram o cálculo das métricas de redes complexas de cada rede por meio da biblioteca *igraph* e R, também um dos objetivos definidos inicialmente que foi atingido, tendo como resultado planilhas contendo todas as métricas.

No entanto, para permitir a posterior análise e reconhecimento de padrões, os arquivos das métricas, gerados separadamente para cada período de avaliação, foram unidos em arquivos únicos, gerando assim arquivos para cada PPGE, Área e Grande Área do conhecimento conforme definidas pela CAPES.

De posse das métricas unificadas, é possível aplicar técnicas de reconhecimento de padrões e obter as características que definem, por exemplo, a pontuação do programa designada pela CAPES, reconhecimento de padrões que corresponde a segunda parte da pesquisa ainda para ser desenvolvida.

## AGRADECIMENTOS

Agradeço ao suporte financeiro providenciado durante a realização da pesquisa pela Universidade Tecnológica Federal do Paraná, Câmpus Cornélio Procópio e também pelo suporte técnico ofertado para a realização das atividades descritas no relatório. Gostaria de agradecer ao Prof. Dr. Fabrício Martins Lopes pelo constante apoio e valiosa orientação no decorrer do projeto.

## REFERÊNCIAS


[1] BARABÁSI, A. *Linked: how everything is connected to everything else and what it means for business, science, and everyday life*. Plume book, 2003.

[2] DIGIAMPETRI, L. A. et al. *BraX-Ray: An X-Ray of the Brazilian Computer Science Graduate Programs.* PloS ONE, San Francisco, Volume 9, 2014.

[3] MENA-CHALCO, J. P., JUNIOR, C., MARCONDES, R. *Brazilian bibliometric coauthorship networks.* Em: Journal of the Association for Information Science and Technology, Voilume 65, pp. 1424-144, 2014.

[4] *The R Project for Statistical Computing*. Disponível em: http://www.r-project.org/. Acesso em 02 de Out. De 2015.

[5] BASTIAN, M., HEYMANN, S. e JACOMY, M. *Gephi: An Open Source Software for Exploring and Manipulating Networks.* Third International AAAI Conference on Weblogs and Social Media, 2009.