COPPE/UFRJ

UMA MÉTRICA PARA RANQUEAMENTO EM REDES DE COLABORAÇÃO BASEADA EM INTENSIDADE DE RELACIONAMENTO

Vinícius Pires de Moura Freire

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Daniel Ratton Figueiredo

Rio de Janeiro
Maio de 2010

UMA MÉTRICA PARA RANQUEAMENTO EM REDES DE COLABORAÇÃO BASEADA EM INTENSIDADE DE RELACIONAMENTO

Vinícius Pires de Moura Freire

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Daniel Ratton Figueiredo, Ph.D.

Prof. Edmundo Albuquerque de Souza e Silva, Ph.D.

Prof.ª Jonice de Oliveira Sampaio, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
MAIO DE 2010

Freire, Vinícius Pires de Moura

Uma métrica para ranqueamento em redes de colaboração baseada em intensidade de relacionamento/Vinícius Pires de Moura Freire. – Rio de Janeiro: UFRJ/COPPE, 2010.

XV, 63 p.: il.; 29, 7cm.

Orientador: Daniel Ratton Figueiredo

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2010.

Referências Bibliográficas: p. 60 – 63.

1. Redes de colaboração. 2. Intensidade de relacionamento. 3. Métrica para ranqueamento. 4. Ranqueamento de programas de pós-graduação. 5. Ranqueamento de pesquisadores. I. Figueiredo, Daniel Ratton. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

À Deus em primeiro lugar. Aos meus familiares e amigos com muito carinho.

Agradecimentos

Primeiramente à Deus pela concessão de sua sagrada luz, proteção e orientação durante todo o mestrado.

À minha família, pelo amor e apoio incondicional em todos os momentos. Em especial aos meus pais que tanto amo, Domingos José e Maria do Carmo (em memória), por toda a educação que me deram ao longo da minha vida. Espero sempre corresponder as suas expectativas e ser um orgulho para vocês. À minha irmã, Marcela, pelo grande amor, carinho e atenção.

Aos meus amigos, Fabrício Raphael e Olivério, por estarem sempre presentes em todas as etapas da minha vida no Rio de Janeiro. Obrigado por me permitirem contar com vocês. Saibam que sempre podem contar comigo, onde estivermos.

Ao meu orientador, Daniel, pela grande atenção dada durante o último ano de mestrado. Sua empolgação e dedicação sempre me deram ânimo para dar o melhor de mim na produção deste trabalho. Sou muito grato por ter tido a oportunidade de tê-lo como orientador. Muito obrigado!!

Ao professores do LAND, por me aceitarem como aluno e confiarem em mim. Obrigado pela atenção e pelos conhecimentos que me passaram ao longo desses anos, tenham certeza que aprendi muito com vocês!

Aos *Satisfactions* pela grande amizade. Obrigado por existirem em minha vida! As lembranças das aventuras que vivemos sempre me darão forças para continuar a minha jornada, pois foram tempos bem aproveitados, inesquecíveis e os mais felizes da minha vida. Luciano, parceiro de idéias, conte comigo, mesmo de longe. Ceará, obrigado por sempre me acompanhar, mesmo nos dias mais puxados de trabalho. Breno, sua amizade é muito importante para mim. Victor, você é o cara! Bruno, o meu amigo mais louco, obrigado!

Aos professores e amigos da UFPI e aos amigos do PoP-PI que me incentivaram a chegar até aqui.

À Janine, João Paulo, Fábio, Larissa, Camila, Sérgio, Augusto, Joice e Eduardo pela amizade desde os tempos da escola.

À Ravena que, mesmo de longe, trouxe muitos dias de alegria com sua animação contagiante.

Ao Renato, por ser meu amigo do peito, amigo de todas as horas.

À Marina Lemos, por ter contribuído de forma significativa para que eu sempre fosse em busca dos meus sonhos e crescesse profissionalmente. Obrigado por ser minha inspiração na superação dos meus limites.

À minha amiga, Marina Meneses, que me deu forças para ultrapassar todas as barreiras e conquistar meus objetivos. Saiba que eu acredito muito em você e torço para que seus sonhos se realizem!

Aos amigos Zezim e Felipe por cederem o apartamento para nós ficarmos quando ainda não tínhamos local para morar.

Ao Hélio, Rodrigo, Heraldo, Jesus, Davi, Orlando, Elenílson e Renan, amigos que fiz na UFRJ e compartilhei ótimos momentos.

Aos amigos do LAND. Ao Fabrício por sempre dar atenção quando eu precisei, sua ajuda foi fundamental no início deste trabalho. Ao Bernardo, Gaspare, Larissa, Luiz, GD, Jefferson, Guto, Guilherme, Marcelo, Leandro, Alejandra, Rafael, Gabriel, Totu e Xandão pela amizade e atenção. À Carol, a mãezona de todos do laboratório! Obrigado pelos cafezinhos e palavras de carinho. "Xandão o Brasil é muito bom!"

À Sukyo Mahikari por me mostrar que posso ser útil à Deus e à sociedade através do meu trabalho.

Aos amigos kumite, em especial à Bruna e Karina, por me incentivarem a sempre colocar Deus em primeiro lugar.

Por fim, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, pelo financiamento deste estudo através da bolsa de estudos, sem a qual seria praticamente impossível a conclusão deste mestrado em uma cidade tão distante de minha terra natal.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

UMA MÉTRICA PARA RANQUEAMENTO EM REDES DE COLABORAÇÃO BASEADA EM INTENSIDADE DE RELACIONAMENTO

Vinícius Pires de Moura Freire

Maio/2010

Orientador: Daniel Ratton Figueiredo

Programa: Engenharia de Sistemas e Computação

Redes sociais vêm sendo estudadas ao longo dos anos em diversas áreas do conhecimento com o objetivo de entender diferentes fenômenos. Redes de colaboração são redes sociais nas quais os relacionamentos representam algum tipo de colaboração profissional entre as pessoas. O estudo de redes de colaboração pode ajudar a identificar indivíduos ou grupos que sejam influentes e importantes dentro daquela comunidade. Intuitivamente, relacionamentos em redes de colaboração possuem diferentes intensidades, que podem ser exploradas para melhor caracterizar um fenômeno qualquer. Este trabalho está dividido em duas partes. A primeira parte constitui um estudo das propriedades topológicas de duas redes de colaboração: a rede de colaboração mundial e a rede de colaboração brasileira de autores de artigos científicos da área de Ciência da Computação. Dentre as propriedades estudadas destacamos a caracterização das intensidades dos relacionamentos destas redes. A segunda parte apresenta uma métrica para ranqueamento de vértice e grupos de vértices baseada na intensidade de relacionamento. Utilizando a métrica proposta e outras métricas clássicas, fazemos um ranqueamento dos programas brasileiros de pós-graduação e dos pesquisadores que atuam no Brasil na área de Ciência da Computação. A avaliação dos resultados foi feita através da comparação com as avaliações subjetivas de programas e pesquisadores feitas pela CAPES e CNPq. Os resultados evidenciam a eficiência da métrica em identificar indivíduos e grupos de indivíduos influentes quando comparada à outras métricas presentes na literatura.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

A METRIC FOR RANKING IN COAUTHORSHIP NETWORKS BASED ON INTENSITY OF RELATIONSHIP

Vinícius Pires de Moura Freire

May/2010

Advisor: Daniel Ratton Figueiredo

Department: Systems Engineering and Computer Science

Social networks have been studied over the years in different areas of knowledge in order to understand various phenomena. Collaboration networks are social networks in which relationships represent some kind of professional collaboration among people. The study of collaboration networks can help identify members or groups that are important and influential within that community. Intuitively, relationships in collaboration networks have different intensities that can be exploited to better characterize phenomenon. This work is divided into two parts. The first part is a study of the topological properties of two collaboration networks, the global collaboration network and the Brazilian collaboration network of authors of scientific papers within the area of Computer Science. Among the properties studied, we focus on the characterization of the intensities of relationships in these networks . The second part presents a ranking metric for vertices and groups of vertices based on the intensities of their relationships. Using the proposed metric and other more classical metrics, we rank the postgraduate Brazilian programs and researchers in Brazil in Computer Science. The evaluation of the proposed metric was performed by comparison with subjective evaluations of researchers and programs made by CAPES and CNPq. The results show the effectiveness of the proposed metric in identifying influential members and groups when compared to another metrics in the literature.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Listagens

Lista de Abreviaturas

BDBComp	Biblioteca Digital Brasileira de Computação, p. 14
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, p. 3
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico, p. 3
DBLP	Digital Bibliography & Library Project, p. 14
LAND	Laboratory for modeling, analysis and development of networks and computer systems, p. 3
PQ	Produtividade em Pesquisa, p. 49
XML	eXtensible Markup Language, p. 16

Capítulo 1

Introdução

Grafos, ou redes, são poderosas ferramentas de abstrações que permitem codificar relacionamentos entre pares de objetos, nos quais vértices representam os objetos e arestas os relacionamentos. Em alguns casos os vértices e as arestas correspondem a objetos físicos do mundo real, em outros, os vértices são objetos reais enquanto as arestas correspondem a relacionamentos intangíveis, e ainda existem casos em que vértices e arestas são puras abstrações [1]. Em redes de transporte, por exemplo, o mapa de rotas utilizado por uma transportadora aérea naturalmente forma um grafo, onde os vértices são os aeroportos, e existe uma aresta entre dois vértices se há um vôo direto entre dois aeroportos. Já em redes de comunicação, um conjunto de computadores conectados através de uma rede de comunicação pode ser modelado como um grafo, onde cada vértice representa um computador e arestas representam conexões físicas entre eles [1].

Dentre os vários tipos de redes, existem as redes sociais. Uma rede social é um conjunto de pessoas ou grupos que possuem algum tipo de relacionamento entre si [2]. Neste caso, relacionamentos entre pessoas podem ser de amizade, de parentesco ou de colaboração (por exemplo, co-autores em um artigo). Em uma rede social de amizade, o relacionamento entre duas pessoas pode representar uma amizade entre elas. Em uma rede de parentesco, relacionamentos entre pessoas podem indicar que as duas pessoas pertencem à mesma família. Os relacionamentos entre grupos de pessoas podem ser de diferentes tipos, por exemplo, um relacionamento entre dois grupos de pesquisa pode representar que estes grupos trabalham ou já trabalharam juntos em algum projeto. Já um relacionamento entre dois times de futebol pode representar que estes times já tenham se enfrentado alguma vez [3] [4].

As características topológicas dessas redes refletem o comportamento social dos seus participantes. Os sociólogos, por exemplo, utilizam-nas exaustivamente para estudar as interações entre pessoas. Elas podem ser utilizadas para identificar a pessoa mais "influente" em uma empresa ou organização e para controlar a propagação de novidades, boatos, piadas, doenças e vírus de e-mail [5].

Intuitivamente, relacionamentos têm diferentes intensidades. Em redes de transporte, por exemplo, a intensidade de uma aresta pode estar relacionada com o número de voos entre os aeroportos. Em redes de comunicação, a intensidade de um relacionamento pode estar relacionada com a quantidade de tráfego que passa pelo canal de comunicação. Então, para medir a intensidade dos relacionamentos, ou peso das arestas, utiliza-se alguma métrica adequada, pois existem diversas maneiras de definir intensidade de relacionamento. Por exemplo, em uma rede de namoro virtual, pode-se medir a intensidade de um relacionamento pelo número de mensagens trocadas entre dois internautas, ou até pelo número de encontros por ano. Neste caso, quanto maior a quantidade de mensagens ou encontros, maior será a intensidade do relacionamento, e consequentemente o peso da aresta. A escolha da métrica para caracterizar a intensidade dos relacionamentos deve ser definida pela aplicação.

Uma rede de colaboração científica é uma rede onde os vértices são os autores de artigos científicos, e existe uma aresta entre dois autores se eles publicaram juntos, ou seja, colaboraram na produção de um artigo científico [6]. É bom enfatizar que redes de colaboração são diferentes de redes de citação, nas quais os nós são documentos e as arestas existem se uma publicação citou a outra. A intensidade dos relacionamentos entre os pesquisadores pode ser medida pelo número total de publicações em conjunto, por exemplo, adicionando peso um à aresta para cada publicação feita por um par de autores. Assim, quanto mais publicações estes dois autores tiverem em conjunto, maior será a intensidade do relacionamento, ou peso da aresta.

Ao analisar uma rede de colaboração pode-se descobrir muitas propriedades topológicas da rede, como o número de autores, o número de publicações, o número de colaboradores por autor, a probabilidade de dois autores terem um colaborador em comum, o menor caminho entre os dois autores mais distantes da rede e o número de componentes conexas. Também é possível identificar outras características importantes que tornam possível o ranqueamento de pesquisadores de acordo com sua importância para um grupo de pesquisa, país ou mundo, ou identificar quais grupos de indivíduos de um país são mais importantes [2] [7] [6].

Logo, a motivação deste trabalho é utilizar redes de colaboração e intensidade de relacionamento para definir importância, seja de indivíduos, de suas relações ou de conjuntos de indivíduos. Assim, através destas métricas, tornar-se possível fazer um ranqueamento de indivíduos ou grupos dentro da rede de colaboração.

1.1 Contribuição

Este trabalho tem duas principais contribuições. A primeira delas é um estudo de diversas propriedades topológicas de duas redes de colaboração, a rede de colaboração mundial e da rede de colaboração brasileira de autores de artigos científicos da área de Ciência da Computação. Neste estudo, destaca-se a caracterização das intensidades dos relacionamentos.

A segunda contribuição é o desenvolvimento de uma métrica baseada em intensidade de relacionamento para medir relevância em redes de colaboração. Este ranqueamento pode ser utilizado para definir a importância ou relevância de indivíduos ou grupos para a rede de colaboração estudada. A validação da métrica é feita através da comparação dos ranqueamentos gerados a partir de métricas objetivas com os ranqueamentos subjetivos feitos por órgãos de grande credibilidade na área acadêmica do Brasil. Dessa forma, os principais pesquisadores e grupos que atuam no Brasil foram identificados por métricas topológicas da rede de colaboração. Os resultados indicam que a métrica proposta tem um desempenho melhor do que outras métricas simples, no sentido de melhor identificar pesquisadores influentes de acordo com uma avaliação subjetiva.

Este trabalho está organizado da seguinte forma. O capítulo 2 aborda o referencial teórico utilizado nesta pesquisa, referencia a caracterização das redes sociais e introduz e justifica algumas métricas utilizadas no decorrer da dissertação.

O capítulo 3 apresenta a caracterização da rede de colaboração estudada. Mostra também como os dados foram coletados e como a rede de colaboração foi construída a partir deles, além de apontar como o conjunto de pesquisadores que atuam no Brasil foi obtido. Também é neste capítulo que é feita a análise das propriedades topológicas da rede mundial e da rede formada somente por pesquisadores que atuam no Brasil. Por fim, apresenta-se uma comparação entre diferentes métricas objetivas.

No capítulo 4 está a principal contribuição deste trabalho. Ele introduz uma métrica para ranqueamento baseado em intensidade de relacionamento e a utiliza para avaliar programas de pós-graduação do Brasil, fazendo uma comparação de seus resultados com a avaliação subjetiva feita pela CAPES e com as métricas já existentes definidas nos capítulos anteriores. O capítulo também utiliza esta métrica para classificar os pesquisadores do país, fazendo uma comparação de seus resultados com a avaliação subjetiva feita pelo CNPq e outra métricas objetivas.

Finalmente, no capítulo 5 são feitas as considerações finais sobre este estudo sugerindo possíveis temas para trabalhos futuros.

Capítulo 2

Trabalhos Relacionados

Neste capítulo são mostrados alguns estudos em redes sociais no contexto de redes de colaboração revisando algumas importantes bibliografias existentes na literatura.

Diversas áreas do conhecimento vêm estudando redes sociais com o objetivo de entender diferentes fenômenos [8] [9]. Por exemplo, na área de saúde, o estudo de redes sociais pode contribuir para o entendimento da propagação de doenças transmissíveis. Na sociologia estudo de redes sociais pode contribuir para a identificação de como novidades se propagam pela sociedade. Na tentativa de entender esses fenômenos, os pesquisadores das diversas áreas vêm desenvolvendo diversas metodologias de análise que tem como base as relações entre os indivíduos, considerando uma estrutura em foma de redes [10].

Redes de colaboração começaram a ser estudadas na Espanha e nos Estados Unidos na tentativa de formar relações de cooperação científica em formato de rede, a partir de indivíduos, grupos e instituições, nacionalmente ou internacionalmente [11].

Uma das redes de colaboração antigas, que ainda hoje é referência na área acadêmica, é a rede de colaboração do grande matemático húngaro Paul Erdős. Através dela, obtém-se o número de Erdős de cada pesquisado [12]. O número de Erdös representa a distância de colaboração entre uma pessoa e Paul Erdös, medida pela autoria de trabalhos acadêmicos.

Para ser atribuído um número de Erdős, um co-autor deve escrever um documento acadêmico com um autor que possua um número finito de Erdős. Paul Erdős é a única pessoa que tem um número de Erdös igual a zero. Para qualquer outro autor, se o menor número de Erdös de todos os seus colaboradores é k, então seu número de Erdös é $k + 1$.

De acordo com [12], Erdös escreveu mais de 1416 artigos científicos, principalmente em colaboração. Ele tinha 504 colaboradores diretos. Estas são as pessoas com número de Erdös igual 1. Ou autores que têm colaborado com eles (mas não com o próprio Erdös) têm um número de Erdös 2 (6593 pessoas), aqueles que têm

colaborado com as pessoas que têm um número de Erdös 2 (mas não com Erdös ou com qualquer um que possua um número de Erdös 1) têm um número de Erdös igual a 3 (33605 autores), e assim por diante. Uma pessoa que não tem um caminho até Erdös na rede de colaboração tem um número de Erdös infinito.

Outra rede que utiliza a mesma idéia é a *The Oracle of Bacon* [13], que determina a distância de um ator qualquer de filmes até Kevin Bacon, sendo que o número de Bacon é 1 se o ator contracenou com Bacon em algum filme.

Em [14], Newman, utilizando redes com pesquisadores das áreas de biologia, física e matemática e procurando responder à uma variedade de questões sobre os padrões de colaboração, encontrou vários resultados através dos estudos dessas redes. Dentre eles, constatou que o numero de colaboradores na rede de pesquisadores da área de biologia é muito maior que na de matemática devido ao modo de pesquisa (biologia trabalha com experimentos em laboratórios com muitas pessoas, e a matemática é mais teórica, trabalhando poucas pessoas em uma pesquisa). Concluiu, também, que nos últimos anos tem crescido o número de colaborações entre os matemáticos devido às mudanças das organizações sociais na comunidade matemática, ao surgimento de melhores sistemas de comunicação, e às possíveis mudanças nos tipos de problemas estudados e abordagens utilizadas.

Em [15], os autores analisaram a produção científica em três regiões diferentes do mundo, Brasil, América do Norte e Europa, por meio de redes de colaboração obtidas a partir de uma base de dados de publicações em Ciência da Computação, a DBLP. Os resultados obtidos por diferentes métricas indicam que o processo de produção do conhecimento tem mudado diferentemente em cada região. A pesquisa é cada vez mais feita em colaboração nas diferentes sub-áreas da Ciência da Computação. O tamanho da componente conexa gigante indica a existência de grupos de colaboração isolados na rede Europeia, ao contrário do grau de conectividade encontrado no Brasil e na América do Norte. Também foi analisada a evolução temporal das redes sociais que representam as três regiões. O número de autores por artigo aumentou em um período de 12 anos. Observou-se que o número de colaborações entre os autores cresce mais rápido que o número de autores.

Em [16], os autores analisaram o crescimento de uma grande rede de colaboração entre pesquisadores da área de Ciência da Computação em um período de 25 anos. Com isso, propuseram um modelo estocástico para predizer eficientemente futuras colaborações entre indivíduos baseada na estrutura da vizinhança local.

Menezes et al. [8] utilizaram um método de detecção de grupos para identificar comunidades de pesquisa na rede social científica brasileira. Os resultados permitiram fazer uma análise detalhada da rede, em especial nos grupos e relacionamentos entre professores. Dentre os aspectos estudados da rede social, pode-se destacar a identificação de áreas interdisciplinares, o nível de cooperação entre instituições e a

identificação de pesquisadores centralizadores do conhecimento.

Freitas et al. [17] apresentaram uma visão geral dos problemas envolvidos na área de descoberta de conhecimento e visualização de informações em redes sociais. O artigo contempla a especificação de um processo para análise de conhecimento e visualização de dados de redes sociais, e discute também aspectos relativos à visualização que permitem de maneira gráfica e interativa explorar as redes sociais.

Na literatura existem diversos estudos de redes sociais utilizando ferramentas para a visualização de redes sociais [17] [18]. A ferramenta Pajek [19], por exemplo, é bastante utilizada para a visualização de grandes redes. Através do Pajek, é possível visualizar a rede de modo recursivo, decompondo-a em estruturas menores e oferecendo ferramentas de análise de estruturas. Além de construir redes com atributos temporais.

Em [6], Newman abordou métricas para medir a intensidade do relacionamento nas redes de colaboração científica. Esta intensidade é representada através de peso nas arestas da rede de colaboração. Primeiramente mostrou-se uma métrica simples que consiste em adicionar peso 1 a uma aresta para cada artigo que um par de autores possui em conjunto. Ou seja, o peso da aresta corresponde ao número de artigos que dois autores escreveram juntos. Um exemplo desta métrica está na ilustrado na figura 2.1, onde existem três artigos escritos pelos autores A, B, C, D e E, cada um com diferentes colaborações.

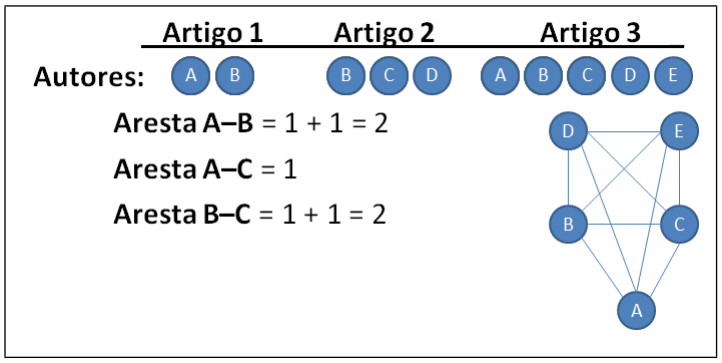


Figura 2.1: Exemplo de cálculo dos pesos das arestas utilizando uma métrica simples.

De acordo com o exemplo da figura 2.1, para calcular o peso da aresta A-B, adiciona-se o peso 1 para cada artigo que eles escreveram juntos. Como eles colaboraram nos artigos 1 e 2, o peso da aresta A-B é 2. Os autores A e C escreveram juntos somente o artigo 2, logo o peso da sua aresta é 1. Já os autores B e C, publicaram juntos os artigos 2 e 3, então o peso da aresta B-C é 2. É importante observar que mesmo o artigo 2 tendo mais autores do que o artigo 3, os pesos das colaborações induzidas por cada artigo tem a mesma importância, gerando os mesmos valores de peso, 1 para cada colaboração.

No mesmo trabalho em questão, Newman introduziu uma nova métrica para

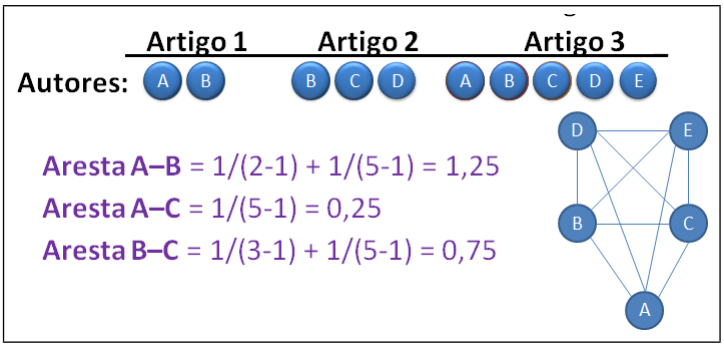


Figura 2.2: Cálculo dos pesos das arestas utilizando o método de Newman.

medir a intensidade do relacionamento em redes de colaboração científicas, daqui para frente chamada de **Métrica de Newman**. Ela funciona da seguinte maneira: cada artigo colaborado por um conjunto de autores adiciona $\frac{1}{n-1}$ à intensidade da colaboração, ou seja, ao peso da aresta, onde n é o número de autores do artigo.

Na figura 2.2, para calcular o peso da aresta A-B utilizando o método de Newman, adiciona-se $\frac{1}{n-1}$ ao peso por cada artigo que os autores A e B escreveram juntos. Eles escreveram sozinhos o artigo 1, logo o peso correspondente à ele é $\frac{1}{2-1} = 1$. O artigo 2 foi escrito por 5 autores, então o peso correspondente à ele é $\frac{1}{5-1} = 0,25$. Logo, o peso desta aresta é 1,25.

É interessante observar os cálculos dos pesos das arestas A-B e B-C, que na métrica simples tiveram o mesmo valor. Já na métrica de Newman, o valor dos pesos foi diferente, pois receberam a influência do número de autores que colaboraram nos artigos. Newman descreve seu método como sendo o quanto os autores se conhecem em cada artigo que eles trabalham juntos. A ideia é que cada artigo tem intensidade constante, que é dividida igualmente entre os co-autores.

Pode-se utilizar a equação 2.1 para calcular o peso da aresta w_{ij} entre os autores i e j:

$$w_{ij} = \sum_k \frac{\delta_i^k \delta_j^k}{n_k - 1} \tag{2.1}$$

onde

$$\delta_i^k = \begin{cases} 1 & \text{se o autor } i \text{ é co-autor do artigo k;} \\ 0 & \text{caso o contrário} \end{cases}$$

e n_k é o número de co-autores do artigo k. Nota-se que artigos com apenas um autor não adiciona peso às arestas.

A figura 2.3 mostra a rede de colaboração correspondente aos dados contidos nas figuras 2.1 e 2.2. Pode-se observar que somando os pesos das arestas que incidem em cada vértice resulta no número de artigos em que o autor é co-autor, ou seja, no número de artigos que o autor possui com algum outro pesquisador. De agora em

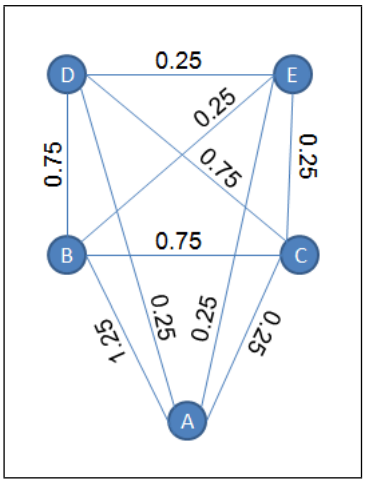


Figura 2.3: Grafo de colaboração dos autores A, B, C, D e E.

diante, essa soma será chamada de peso do vértice e será denotada por p_i:

$$p_i = \sum_{j(\neq i)} w_{ij} = \sum_k \sum_{j(\neq i)} \frac{\delta_i^k \delta_j^k}{n_k - 1} = \sum_k \delta_i^k \tag{2.2}$$

As equações 2.1 e 2.2 foram utilizadas neste trabalho para calcular, respectivamente, os pesos das arestas e dos vértices.

2.1 Coeficiente de Gini

Dentre os conceitos utilizados para desenvolver esta dissertação estão a curva de Lorenz e coeficiente de Gini. Estes conceitos são bastante empregados na área de Economia. Eles são utilizados para calcular a dispersão de uma distribuição empírica de uma determinada característica de uma população. Em geral, são utilizados para determinar o nível de desigualdade na distribuição de renda de um país.

A curva de Lorenz foi introduzida pelo economista americano Max Otto Lorenz em 1905 como uma maneira de comparar facilmente a desigualdade entre populações de tamanhos ou níveis de renda diferentes, permitindo comparações da situação de um país ao longo dos anos ou comparações entre países [20].

Em geral, em uma curva de Lorenz, tem-se o eixo X como sendo a porcentagem acumulada de pessoas de uma região ou país em ordem crescente e o eixo Y sendo a porcentagem de renda acumulada das pessoas.

A Figura 2.4 mostra um exemplo desta curva. Uma distribuição igualitária, onde cada indivíduo tem a mesma renda, é representada no gráfico por uma reta entre os pontos [(0,0) e (1,1)] [21]. Ou seja, x% da população detém x% da renda. Porém, a distribuição de renda real é desigual, sendo mais parecida com a do país A ilustrado na figura. À medida que a curva de Lorenz se afasta da reta, da igualdade perfeita,

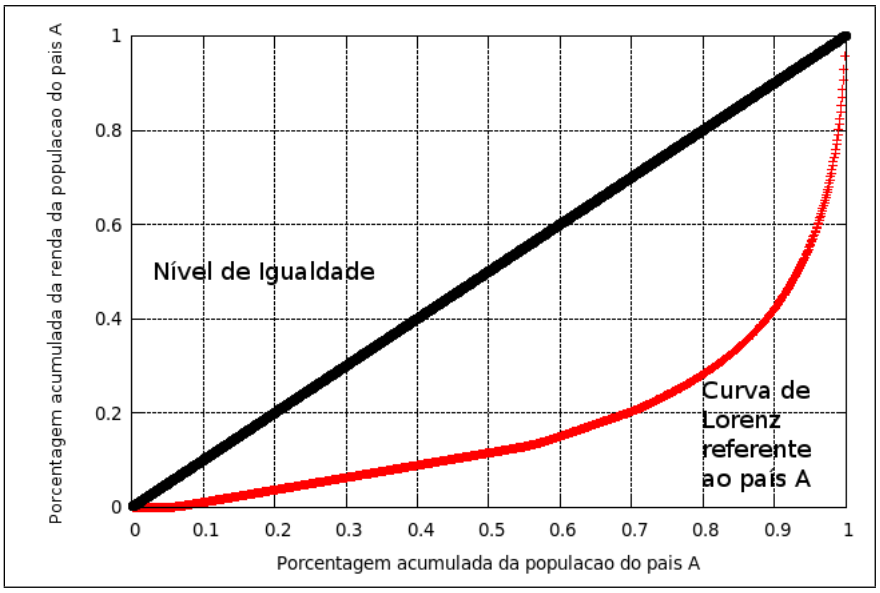


Figura 2.4: Distribuição dos rendimentos da população do país A.

o nível de desigualdade aumenta.

Em [20], Marcelo Medeiros descreve uma analogia interessante. Ele compara toda a renda de uma sociedade a um grande bolo distribuído para pessoas que participam de um desfile. Durante a marcha, cada pessoa recebe uma fatia do bolo proporcional à sua renda. Ordenam-se estas pessoas de acordo com suas rendas em ordem crescente. Logo, os primeiros a marchar receberão pequenos pedaços do bolo e, os últimos, pedaços bem maiores. A qualquer instante, é possível verificar a quantidade de bolo que ainda resta. Desta forma obtém-se o quanto do bolo foi distribuído para os x% mais pobres da população. Vendo o gráfico de forma contrária, se 40% do bolo foi distribuído para 80% da população, é porque os outros 60% estarão reservados para os 20% mais ricos. Marcando no gráfico qual a proporção de pessoas que já marcharam e a proporção do quanto foi distribuído do bolo tem-se uma curva de Lorenz.

Feita a analogia, fica fácil entender a figura 2.4. Observando o ponto (0,9;0,42), nota-se que 90% da população mais pobre detém apenas 42% da renda do país A, e somente 10% dos mais ricos possuem o grande montante restante de 58% de toda a renda do país. Desta forma, é possível ver a grande desigualdade desta distribuição. Porém, existe uma necessidade de expressar essa desigualdade em apenas um número, facilitando a comparação entre diferentes distribuições. Para isto, surgiu o coeficiente de Gini.

O coeficiente de Gini foi proposto pelo matemático italiano Conrad Gini e é uma medida internacional de desigualdade da distribuição de renda. Esta é calculada dividindo-se a área entre a reta da desigualdade perfeita e a curva de Lorenz de um determinado país, pela área do triângulo formado pela reta, o eixo horizontal e

o eixo vertical do lado direito da figura 2.4. Em outras palavras, é calculado pela equação

$$G = \frac{A}{A+B} \tag{2.3}$$

onde A e B são as áreas representadas na Figura 2.5, o eixo x representa a porcentagem acumulada da população e o eixo y representa a porcentagem acumulada da renda da população.

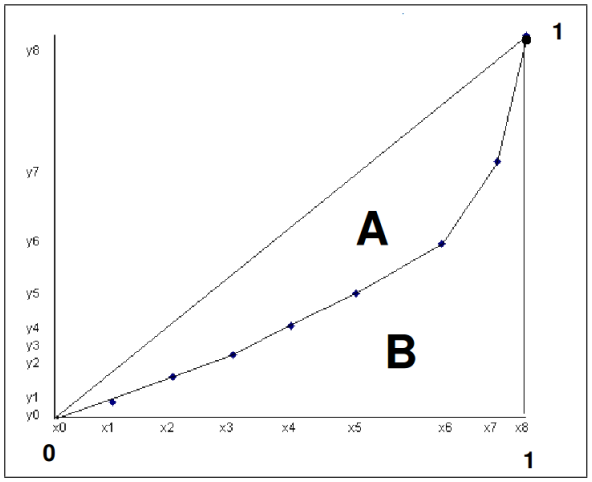


Figura 2.5: Representação gráfica do coeficiente de Gini.

Para demonstrar como calcular o coeficiente de Gini, a equação será desenvolvida:

$$G = \frac{A}{A+B} = \frac{A+B}{A+B} - \frac{B}{A+B} = 1 - \frac{B}{A+B} \tag{2.4}$$

Como $A+B$ é igual à área do triângulo de base 1 e altura 1, $A + B = \frac{1}{2}$, logo

$$G = 1 - 2B \tag{2.5}$$

Dividindo a área B em n trapézios, de acordo com a figura 2.6 e lembrando que a área do trapézio é dada por

$$T = \frac{(b_1 + b_2) \times h}{2} \tag{2.6}$$

onde b_1 e b_2 são as bases do trapézio e h a altura, pode-se calcular a área B somando-se as áreas de todos os trapézios [22]. Observando que a área do trapézio delimitado pelos pontos (x_1, y_1) e (x_2, y_2) é calcula por

$$T = \frac{(y_1 + y_2) \times (x_2 - x_1)}{2} \tag{2.7}$$

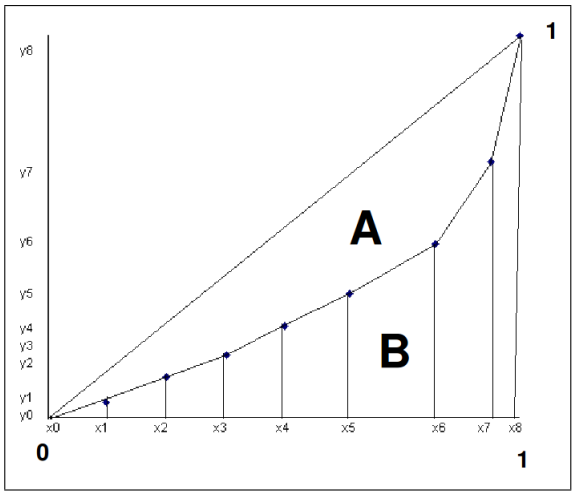


Figura 2.6: Representação gráfica do coeficiente de Gini e divisão da área B em n trapézios.

Tem-se então que o valor da área B é dado por:

$$B = \sum_{k=1}^{n-1} \frac{(y_k + y_{k+1}) \times (x_{k+1} - x_k)}{2} \tag{2.8}$$

Substituindo 2.8 em 2.5, tem-se

$$G = 1 - 2 \times \left[\sum_{k=1}^{n-1} \frac{(y_k + y_{k+1}) \times (x_{k+1} - x_k)}{2} \right] \tag{2.9}$$

O coeficiente de Gini é dado, então, por

$$G = 1 - \sum_{k=1}^{n-1} (y_k + y_{k+1}) \times (x_{k+1} - x_k) \tag{2.10}$$

O valor de G varia entre 0 e 1 e quanto mais próximo de 0, mais igual é a distribuição de renda ou riqueza, enquanto um elevado coeficiente de Gini indica a distribuição mais desigual. O valor zero corresponde à perfeita igualdade (todos tem exatamente a mesma renda) e 1 corresponde à desigualdade perfeita (na qual uma só pessoa possui toda a renda, enquanto as outras pessoas não possuem renda alguma). De acordo com os dados em [23], o coeficiente de Gini no mundo varia, aproximadamente, de 0.247 na Dinamarca a 0.743 na Namíbia. Já no Brasil, com um valor de 0.55, o coeficiente de Gini confirma a grande desigualdade de renda do Brasil. A figura 2.7 mostra os coeficientes de Gini na maioria dos países do mundo.

Neste trabalho utilizamos o coeficiente de Gini para medir a desigualdade de distribuições empíricas sobre a intensidade dos relacionamentos de uma determinada população.

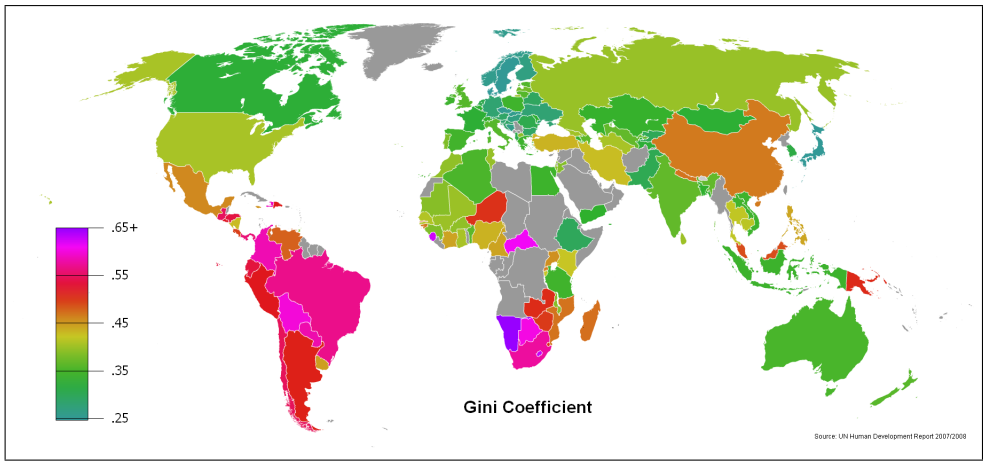


Figura 2.7: Mapa-múndi do coeficiente de Gini (2007/2008) [24].

2.2 Precisão, Abrangência e Medida-F

As métricas precisão, abrangência e medida-F são medidas frequentemente utilizadas na área de recuperação da informação para avaliar os resultados de uma busca ou ranqueamento. Para esta avaliação ocorrer é necessário haver uma comparação dos resultados com os dados corretos, e assim, detectar quão bons são os métodos utilizados no ranqueamento.

Duas métricas são utilizadas para fazer esta avaliação: precisão e abrangência.

Precisão é a proporção de um conjunto de objetos retornados que é realmente relevante [25]. Abrangência é a proporção de objetos relevantes que foram retornados [26].

A figura 2.8será utilizada para explicar os conceitos de precisão e abrangência.

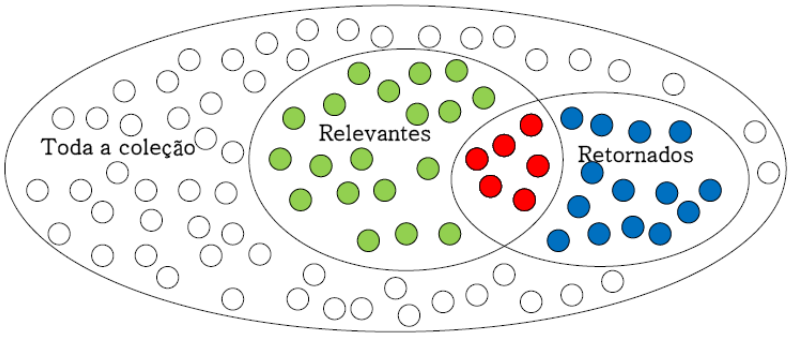


Figura 2.8: Toda a coleção de documentos com destaque para os recuperados e relevantes.

As medidas precisão e abrangência são definidas por:

$$\text{Precisão} = \frac{\text{Número de objetos relevantes retornados}}{\text{Número total de objetos retornados}} \tag{2.11}$$

$$\text{Abrangência} = \frac{\text{Número de documentos relevantes retornados}}{\text{Número total de documentos relevantes}} \tag{2.12}$$

Ou utilizando as cores dos objetos da figura 2.8 para a definição, pode-se definir precisão e abrangência por:

$$\text{Precisão} = \frac{\text{Vermelhos}}{\text{Azuis} + \text{Vermelhos}} \tag{2.13}$$

$$\text{Abrangência} = \frac{\text{Vermelhos}}{\text{Verdes} + \text{Vermelhos}} \tag{2.14}$$

Os dois valores devem ser sempre calculados para um determinado conjunto de objetos retornados e estão compreendidos entre zero e um. Um cenário ideal seria ter sempre uma precisão e abrangência igual a um, o que significa que todos e apenas os objetos relevantes são retornados. Entretanto, em um sistema real, ao melhorar uma das medidas em geral deteriora a outra. Para melhorar a precisão, deve-se diminuir o número de objetos retornados, porém, isso diminui a abrangência. Para melhorar a abrangência, deve-se aumentar a quantidade de objetos retornados. Porém a precisão irá diminuir.

Este compromisso entre as duas métricas torna difícil a qualificação da qualidade de um conjunto de resultados. A medida-F facilita essa análise, pois utiliza apenas um valor numérico entre 0 e 1. A medida-F identifica situações em que os resultados contém informações desnecessárias (baixa precisão), e quando os resultados não contêm informação suficiente (baixa abrangência) [27].

A medida-F é uma média harmônica que considera ao mesmo tempo a precisão e abrangência. Esta métrica é dada pela equação 2.15.

$$F = \frac{2 \times \text{ Precisão } \times \text{ Abrangência}}{\text{Precisão } + \text{Abrangência}} \tag{2.15}$$

Capítulo 3

Caracterização da Rede de Colaboração

Este capítulo apresenta os dados que foram utilizados para criar a rede de colaboração que foi estudada e descreve como o conjunto de pesquisadores que atuam no Brasil foi obtido. Além disto, apresenta uma análise das diferentes propriedades topológicas da estrutura da rede mundial e da rede formada somente por pesquisadores que atuam no Brasil. Por fim, estabelece uma comparação direta entre estas duas redes.

3.1 Dados

A rede de colaboração científica foi construída utilizando bases de dados disponíveis publicamente na Web contendo informações acerca das publicações na área da Ciência da Computação.

Durante a busca, algumas bases foram identificadas, dentre elas a plataforma Lattes do CNPq, a DBLP e o BDBComp. A plataforma Lattes é uma base de dados que contém currículos e instituições de todas as áreas do conhecimento com cerca de 1.620.000 currículos (8% doutores e 13% mestres) [28]. É uma base brasileira e referência nacional no meio acadêmico com a maior parte de seus usuários sendo brasileiros. O conteúdo é inserido pelo próprio dono do currículo, e assim, não há um padrão na escrita dos dados e, consequentemente, existem duplicidades de identidade de autores e artigos. Esse foi um dos motivos por não se ter utilizado a plataforma Lattes na implementação deste trabalho. Outros motivos foram o pequeno número de pesquisadores estrangeiros cadastrados na plataforma, bem como a dificuldade de coletar todos os dados, pois a base completa não está publicamente disponível.

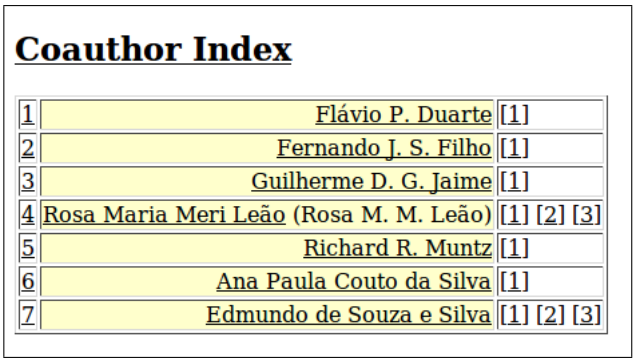
A base BDBComp (Biblioteca Digital Brasileira de Computação) é uma base que contém trabalhos publicados em periódicos nacionais e anais de eventos realizados

no Brasil na área de Ciência da Computação [?]. Esta base não foi utilizada na construção da rede de colaboração pois não contém trabalhos publicados fora do Brasil. Outra base analisada foi a *DBLP* (*Digital Bibliography & Library Project*) [29]. A *DBLP* é uma base de dados com informações bibliográficas dos principais periódicos e conferências da área de Ciência da Computação, com cerca de 1,3 milhões de publicações e 720.000 autores. É referência mundial no meio acadêmico da computação e muito utilizada por pesquisadores do meio para coletar detalhes bibliográficos ao compor as listas de referências para novos artigos [30]. Diferente do Lattes, seu sítio na *Web* não mostra o currículo dos autores, mas sim todas as publicações cadastradas de cada pessoa, bem como todos os seus co-autores. Também diferente do Lattes, os autores não podem cadastrar nenhuma informação diretamente. As figuras 3.1 e 3.2 ilustram o *website* da DBLP, disponível em [29]. É uma base que também possui duplicidades de autores, porém em bem menor quantidade, pois é mantida apenas pelo seu idealizador, o professor Michael Ley, da Universidade de Trier, Alemanha.

Figura 3.1: Publicações de um autor no sítio da *DBLP*.

A *DBLP* foi a base de dados utilizada nesta dissertação. Dentre os motivos que levaram à sua escolha estão:

- Contém principalmente publicações e autores que estão relacionados com a Ciência da Computação, dando origem a um conjunto de pessoas mais coeso;
- Contém autores do mundo inteiro;
- Está em constante atualização;



Coauthor Index

1	Flávio P. Duarte	[1]
2	Fernando J. S. Filho	[1]
3	Guilherme D. G. Jaime	[1]
4	Rosa Maria Meri Leão (Rosa M. M. Leão)	[1] [2] [3]
5	Richard R. Muntz	[1]
6	Ana Paula Couto da Silva	[1]
7	Edmundo de Souza e Silva	[1] [2] [3]

Figura 3.2: Co-autores do autor da Figura 3.1 em página da Web no sítio da *DBLP*.

- Encontra-se completamente disponível publicamente na Web em formato XML;

Uma vez obtida a base de dados, é necessário construir a rede de colaboração utilizando os dados da base. O arquivo obtido, *dblp.xml*, está em formato XML e organizado como no exemplo da Listagem 3.1.

Listagem 3.1: Exemplo do arquivo dblp.xml.

```
<www mdate="2009-01-29" key="homepages/l/RosaMMLeao">
<author>Rosa Maria Meri Le&atilde;o</author>
<author>Rosa M. M. Le&atilde;o</author>
<title>Home Page</title>
</www>

<inproceedings mdate="2009-04-15" key="conf/icc/RochaLS07">
<author>Antonio Augusto de Arag&atilde;o Rocha</author>
<author>Rosa Maria Meri Le&atilde;o</author>
<author>Edmundo de Souza e Silva</author>
<title>An End-to-End Technique to Estimate the Transmission Rate of an
    IEEE 802.11 WLAN.</title>
<pages>415-420</pages>
<year>2007</year>
<booktitle>ICC</booktitle>
<ee>http://dx.doi.org/10.1109/ICC.2007.75</ee>
<crossref>conf/icc/2007</crossref>
<url>db/conf/icc/icc2007.html#RochaLS07</url>
</inproceedings>
```

A primeira parte do código (linhas 1-5) da listagem 3.1, é um exemplo de um trecho com informações de um autor cadastrado na *DBLP*. Pode-se observar que a autora em questão tem dois nomes cadastrados na base (Rosa Maria Meri Leão e Rosa M. M. Leão), então, independente da forma de seu nome, ao aparecer em uma publicação, estes estarão referenciando a mesma pessoa.

A segunda parte do código (linhas 7-18) é um exemplo de um trecho do arquivo

contendo informações de uma publicação cadastrada na *DBLP*. Cada *tag* delimita um tipo de informação, por exemplo, entre as *tags* <*author*> e </*author*> está contido um autor do artigo. Como elas aparecem três vezes, isso indica que existem três autores nesta publicação. Os pares de *tags* <*title*>, </*title*> e <*year*>, </*year*> referem-se, respectivamente, ao título e ano da publicação.

A partir do arquivo *XML* é possível extrair todos os dados fazendo um *parser*. Em [31] apresenta um exemplo de *parser* para o arquivo *dblp.xml*. Para construir a rede de colaboração científica utilizado nesta dissertação foi implementado um *parser* em *java* semelhante devido à necessidade de adaptação do mesmo para introduzir a noção de peso. Na rede de colaboração construída, cada vértice é um autor cadastrado na DBLP e existe uma aresta entre um par de autores se eles são co-autores em ao menos uma publicação. As arestas contêm pesos calculados através da métrica de Newman apresentada no capítulo 2. Vale ressaltar que todos os nomes de um mesmo autor são mapeados para um mesmo vértice da rede.

Conforme apresentado no capítulo 1, um dos objetivos é avaliar e comparar grupos e pesquisadores que atuam no Brasil. Para isto, é necessário identificar a nacionalidade ou local de trabalho dos vértices da rede de colaboração.

A *DBLP* não classifica os autores por nacionalidade nem por local de trabalho, então foi necessário procurar um método para obter um conjunto de autores que atuam no Brasil. A base de dados DBLP, na lista de autores, pode apresentar o endereço de sua página pessoal, como no exemplo da listagem 3.2. Então, de início, todos os autores com página pessoal contendo .br/ foram considerados como atuantes no Brasil. Porém a quantidade de autores com esta característica foi muito pequena, pouco mais de duzentos. Foi preciso encontrar outra forma de identificar mais autores que atuassem no Brasil.

Listagem 3.2: Exemplo de autor com página pessoal cadastrada.

```
<www mdate="2004-03-31" key="homepages/a/VirgilioAlmeida">
<author>Virg&iacute;lio A. F. Almeida</author>
<author>Virgilio Almeida</author>
<title>Home Page</title>
<url>http://www.dcc.ufmg.br/~virgilio/</url>
</www>
```

Para aumentar essa lista, buscou-se os pesquisadores dos programas de pós-graduação das universidades brasileiras disponíveis em [32]. Muitos dos nomes contidos em [32] estão cadastrados de forma diferente na *DBLP*, por exemplo, abreviados, dificultando a identificação dos autores. Para agilizar a busca, desenvolveu-se um algoritmo que compara os nomes dos pesquisadores com os autores da *DBLP* e lista os nomes prováveis de serem os pesquisadores desejados. Porém, como na maioria das vezes os nomes na *DBLP* estão abreviados, a escolha do nome correto

na lista dos prováveis nomes foi feita manualmente, comparando a página do autor no sítio da *DBLP* com seu currículo Lattes, para saber se os diferentes nomes representavam a mesma pessoa.

Repetiu-se o mesmo procedimento utilizando a lista de pesquisadores com bolsa de produtividade em pesquisa do CNPq, coletados a partir de [33]. Com estas duas listas, o tamanho do conjunto dos pesquisadores que atuam no Brasil cresceu de 200 para 850. Porém, estes autores têm fortes relações com outros profissionais brasileiros (ou que atuam no Brasil) e com seus alunos. Estas pessoas deveriam ser acrescentados à lista de atuantes no Brasil.

Para identificar potenciais colaboradores de pessoas já identificadas como atuantes no Brasil, analisou-se os vértices vizinhos ao conjunto dos brasileiros já selecionados e, manualmente, verificou-se se o autor atuava no Brasil. Para reduzir o número de vértices a serem inspecionados manualmente, ordenou-se os mesmos pela intensidade que eles estavam relacionados com o conjunto primeiramente identificado como atuantes no Brasil. Para cada autor verificado, quando este atuava no Brasil, era adicionado ao conjunto de brasileiros. Esse processo foi repetido diversas vezes, até que o peso das arestas entre os membros do conjunto e os autores de fora que poderiam atuar no Brasil (desconsiderando os estrangeiros já identificados) fosse muito pequeno. Assim, o conjunto de pessoas que atuam no Brasil ficou com exatamente 2.729 pessoas.

No entanto, apesar do conjunto acima possuir um bom número de autores que trabalham no Brasil, existem autores que não foram considerados. Isso acontece devido ao processo manual de classificação, o qual leva muito tempo. Porém, estas imprecisões não afetam de forma significativa os estudos realizados nesta dissertação, pois uma grande massa de pesquisadores atuantes no Brasil foi identificada corretamente. Em particular, os pesquisadores de maior intensidade (peso) foram todos considerados.

Uma vez coletados estes dados, construída a rede de colaboração e identificado o conjunto de pessoas atuante no Brasil, a estrutura topológica da rede foi analisada utilizando várias métricas. Nas próximas seções serão apresentadas a análise da rede como um todo e da rede formada apenas pelo conjunto de pesquisadores que atuam no Brasil.

3.2 Análise da Rede Mundial

De início foi feita a análise das propriedades estruturais da rede de colaboração mundial, construída através dos dados da DBLP. Cada autor cadastrado corresponde a um vértice e as arestas correspondem à colaboração entre os autores. A rede de colaboração mundial possui 722.392 vértices e 2.272.540 arestas. As métricas foram

obtidas utilizando uma biblioteca para manipulação de grafos desenvolvida no LAND [?], a Libgraph. Entretanto, diversas novas funções foram adicionadas à biblioteca na execução dessa dissertação, como por exemplo, a métrica para ranqueamento baseada em intensidade de relacionamento.

Para definir formalmente as métricas avaliadas, a rede de colaboração foi representada por um grafo não direcionado $G = \{V, E\}$, no qual os autores correspondem ao conjunto de vértices $V = \{v_i\}$ e as colaborações ao conjunto de arestas representado por $E = \{e_{ij}\}$, onde $e_{ij} = (vi, vj)$.

3.2.1 Grau

O grau d_i é o número de vizinhos do vértice i, ou seja, o número de colaboradores do autor i. O número de vizinhos é muito importante para a caracterização da rede, pois se, por um lado, um vértice não tem vizinhos, este não pode trocar informações com outros vértices e assim, não tem influência alguma na rede. Ou, por outro lado, se um vértice possui muitos vizinhos, pode trocar informações com qualquer um deles e possuir significativa intensidade de relacionamento. Matematicamente, o grau de um vértice é dado por [34]:

$$d_i = \sum_{v_j \in V} e_{ij} \tag{3.1}$$

Para caracterizar a estrutura da rede, analisou-se primeiramente a distribuição empírica do grau.

A distribuição empírica do grau dos vértices é dada por

$$f_d(k) = \frac{\text{Número de vértices com grau } k}{\text{Número total de vértices}} \tag{3.2}$$

onde $f_d(k)$ é a fração relativa de vértices com grau igual a k.

No entanto, a distribuição do grau é melhor representada utilizando sua função complementar cumulativa:

$$P[D \geq k] = \sum_{k'=k}^{\infty} f_d(k') \tag{3.3}$$

onde $P[D \geq k]$ é a fração relativa de vértices com grau maior ou igual k [2].

A distribuição empírica do grau é apresentada na figura 3.3, onde o eixo x representa o grau dos vértices, ou seja, o número de colaborações dos autores e o eixo y é a fração dos vértices com grau maior ou igual a x, ou seja, a fração dos autores com número de colaborações maiores ou iguais a x.

Observando a figura 3.3 pode-se perceber que a distribuição do grau possui uma

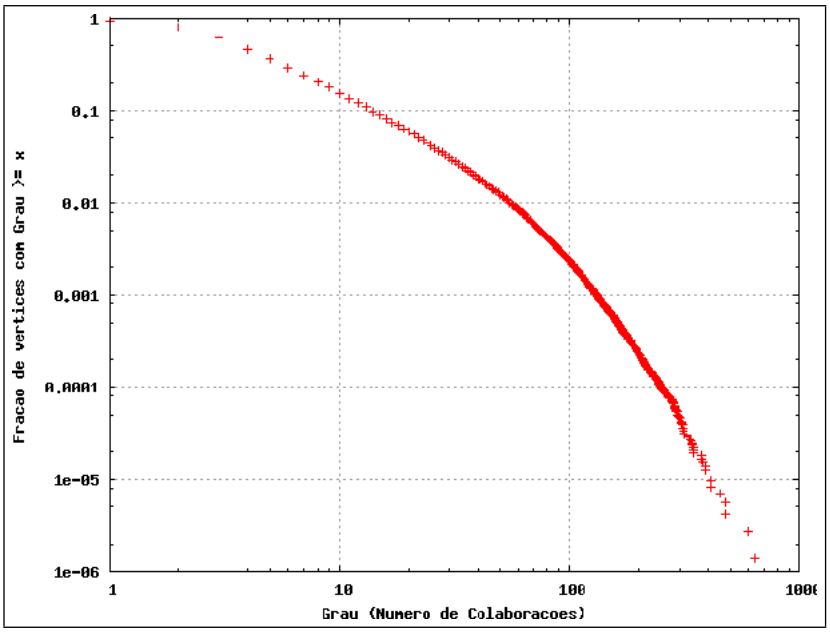


Figura 3.3: Distribuição do grau dos vértices da rede de colaboração mundial.

cauda pesada, ocorrendo valores muito distantes do grau médio, que é 6.3, variando de 0 a 643. Há um pequeno número de autores com muitas colaborações e uma grande quantidade de autores com poucos colaboradores, por exemplo, no gráfico observa-se que 15% dos vértices têm grau maior ou igual 10 e que 80% dos vértices têm grau menor que 8.

Através da distribuição do grau, nota-se também uma característica não esperada: 6% dos autores não possuem nenhuma colaboração, correspondendo a mais de 43.000 autores. Estes autores não possuem nenhuma aresta na rede de colaboração[1].

3.2.2 Componentes Conexas

Um grafo é chamado de conexo se existe um caminho entre qualquer par de vértices. Definindo formalmente, um grafo $G = (V, E)$ é conexo se, dados dois nós v_1, v_2, existe uma cadeia de vizinhos com extremidade inicial v_1 e extremidade final v_2 [1]. A figura 3.4 é um exemplo de grafo conexo. Para verificar se um determinado grafo é conexo basta utilizar a Busca em Largura (*BFS - Breadth First Search*) e para verificar se todos os vértices foram visitados.

Em redes de colaboração, as idéias dos autores de uma rede conectada podem chegar aos demais através de um ou múltiplos saltos. Quando o grafo não é conexo, podem existir vários subgrafos conexos, surgindo, assim, a ideia de componentes conexas. As componentes conexas são os maiores subgrafos conectados de um grafo, mais precisamente, são os maiores conjuntos de nós, tal que todos os nós conseguem alcançar os demais. A figura 3.5 mostra um exemplo de um grafo desconexo e a figura

[1] O ponto $x = 0$ não aparece na figura, pois o gráfico está em escala log-log.

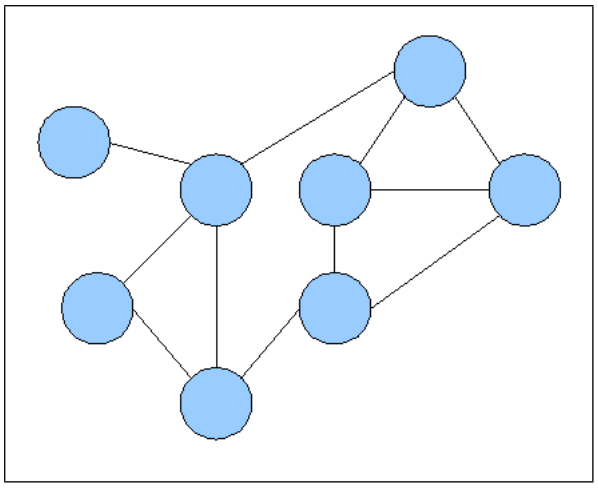


Figura 3.4: Exemplo de grafo conexo.

3.6 apresenta separadamente as componentes conexas identificadas por diferentes cores e, pode-se observar que um vértice isolado (que não possui arestas conectando a outros nós) também é uma componente conexa, pois alcança a si mesmo [34].

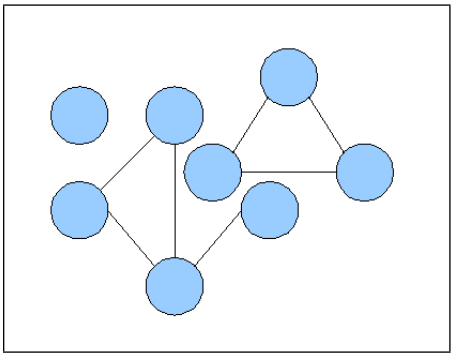


Figura 3.5: Exemplo de um grafo desconectado.

A distribuição empírica do tamanho das componentes conexas é apresentada na figura 3.7, onde o eixo x representa o tamanho da componente conexa e o eixo y é a fração das componentes conexas com tamanho maior ou igual a x. Através da figura 3.7 pode-se constatar que existe uma componente gigante e várias outras muito pequenas. Na componente gigante estão quase todos os vértices do grafo, 576.309 vértices, constituindo 79,8% dos vértices da rede de colaboração, e a segunda maior componente é pequena, com apenas 42 vértices.

As componentes conexas da rede utilizada neste trabalho têm, em média, o tamanho igual a 9,3 e no total são 77.493 componentes. Como é um grafo com muitos nós (722.392), pode-se concluir que existem muitas componentes com poucos vértices. Para esclarecer, o gráfico da distribuição do tamanho das componentes conexas é apresentado na figura 3.7. Observa-se, então, que apenas 0,8% das componentes conexas são maiores ou iguais a 9.

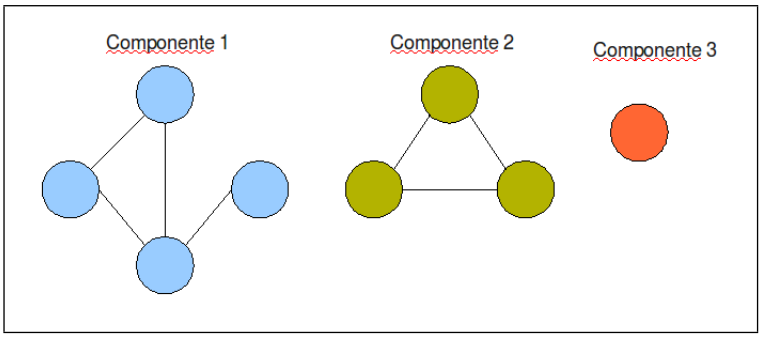


Figura 3.6: Componentes conexas do grafo apresentado na figura 3.5.

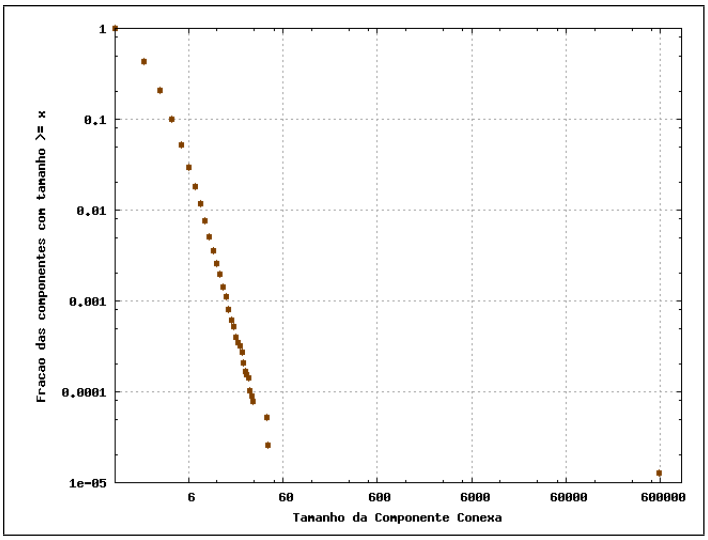


Figura 3.7: Distribuição do tamanho das componentes conexas da rede de colaboração mundial.

3.2.3 Coeficiente de Clusterização

O coeficiente de clusterização C_i de um vértice i mede a conectividade entre os vizinhos de i, por exemplo, se um vértice A está relacionado com B e C e deseja-se calcular a probabilidade dos vértices B e C estarem relacionados, utiliza-se o coeficiente de clusterização [2].

O cálculo C_i é dado pela razão do número de arestas entre os vizinhos de i com o número máximo de possíveis arestas entre os vizinhos de i [35]. A fórmula de C_i é apresentada pela equação 3.4.

$$C_i = \frac{E_i}{d_i \times (d_i - 1)/2} \tag{3.4}$$

A figura 3.8 apresenta um exemplo de cálculo de C_i. Nela, o nó i tem três vizinhos, que podem ter no máximo de 3 conexões entre eles. No primeiro grafo, as três conexões possíveis são realizadas (linhas pretas grossas), resultando em um coeficiente de clusterização igual a 1. Na parte central da figura, apenas uma conexão é realizada (linha preta) e 2 conexões estão faltando (linhas tracejadas vermelhas), re-

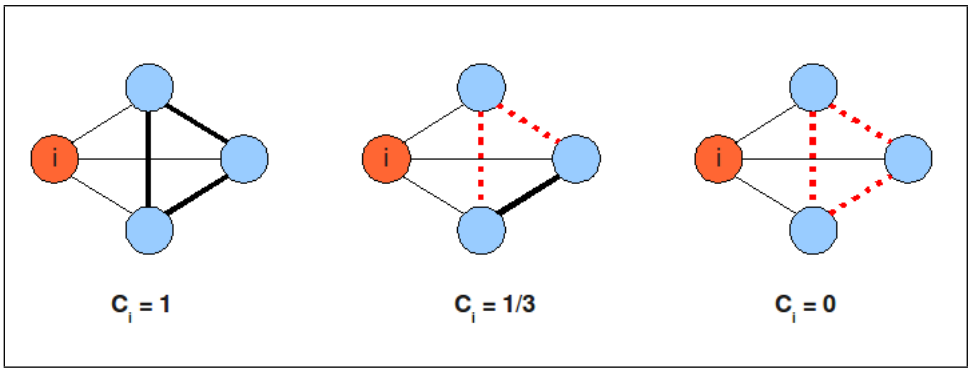


Figura 3.8: Exemplo de cálculo do coeficiente de clusterização.

sultando em $C_i = \frac{1}{3}$. Finalmente, no terceiro grafo, nenhuma das possíveis conexões entre os vizinhos do nó i são realizadas, produzindo um coeficiente de clusterização igual a 0.

Para calcular o coeficiente de clusterização do grafo, basta fazer a média aritmética dos coeficientes de todos os vértices. Na rede de colaboração em questão, o coeficiente de clusterização é 0,59. Logo, a chance de existir uma colaboração entre dois autores que possuem um colaborador em comum é, em média, relativamente alta. A distribuição do coeficiente de clusterização está na figura 3.9, a partir dela, observa-se que 46% dos vértices tem coeficiente de clusterização entre 0,9 e 1,0 e que apenas 25% dos vértices tem um coeficiente de clusterização mais baixo que 0,1.

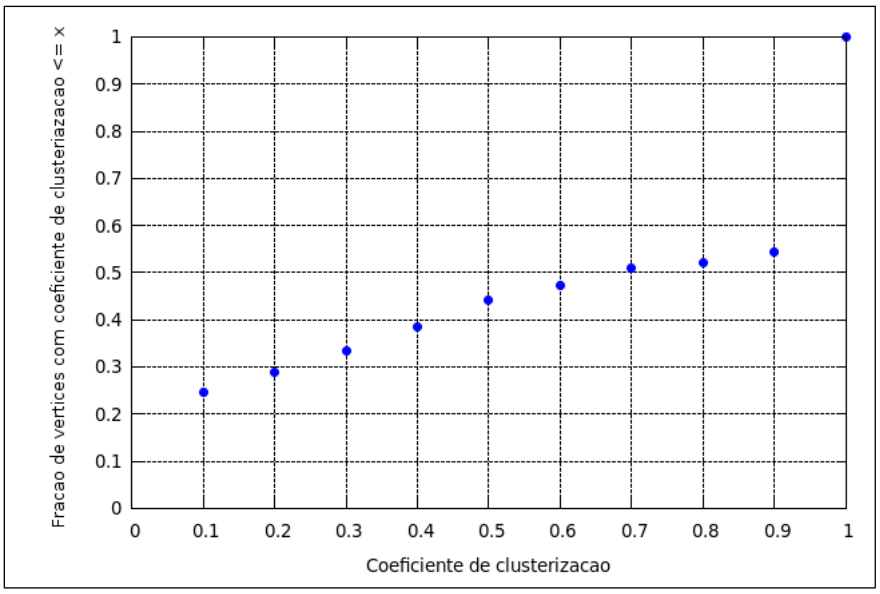


Figura 3.9: Distribuição do coeficiente de clusterização da rede de colaboração mundial.

3.2.4 Distância

A distância entre um par de vértices é dada pelo número de arestas no menor caminho entre eles [35]. Ela pode ser representada através de uma função $d(v_1, v_2)$, onde v_1 e v_2 são vértices. Quando não há caminho entre v_1 e v_2, $d(v_1, v_2) = \infty$. Para calcular a distância média do grafo, basta fazer a média aritmética das distâncias entre todos os pares de vértices do grafo, como ilustrado na equação 3.5.

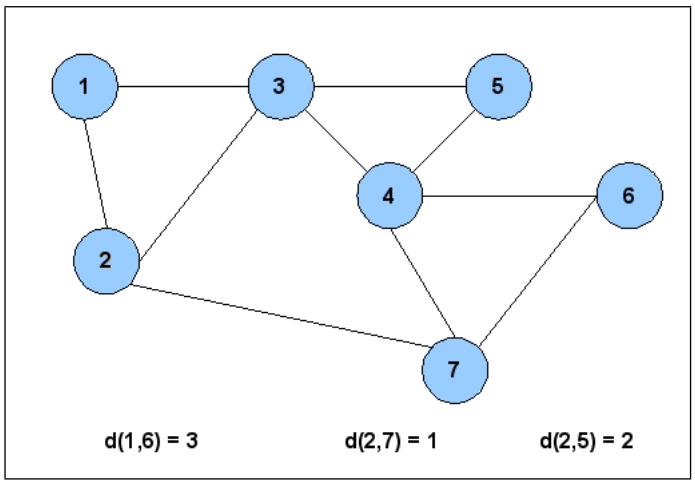


Figura 3.10: Exemplo de cálculo da distância entre os pares de vértices (1,6), (2,7) e (2,5).

Como exemplo do cálculo de distância, apresenta-se a figura 3.10. Através desta métrica, é possível determinar a quantos graus de separação estão a maioria dos autores. A rede estudada possui uma distância média de 6,3. É interessante observar que ela possui o efeito "mundo pequeno", pois mesmo tendo um grande número de vértices, a distância média é pequena, ou seja, é proporcional ao logaritmo do número de vértices. Ao mesmo tempo a rede possui um alto grau de clusterização, possuindo muitos triângulos. O efeito "mundo pequeno" e o conceito de "seis graus de separação" foi identificado pelo psicólogo-social Stanley Milgram (1967), que realizou experimentos com pessoas nos Estados Unidos e identificou caminhos curtos entre as pessoas, com média em torno de 5 saltos [35] [36].

$$\bar{d} = \frac{\sum_{v_1, v_2 \in V} d(v_1, v_2)}{\binom{n}{2}} \tag{3.5}$$

A distribuição da distância é apresentada na figura 3.11. Através dela fica evidente sua média e também observa-se que mais de 75% dos pares de vértices tem distância 5, 6 ou 7. Além disso, ela mostra que distâncias bem maiores existem, mas são muito poco frequentes. Logo, o diâmetro da rede, ou seja, o maior caminho mínimo entre dois vértices no grafo, dado por r na equação 3.6, é 23 na rede de colaboração mundial.

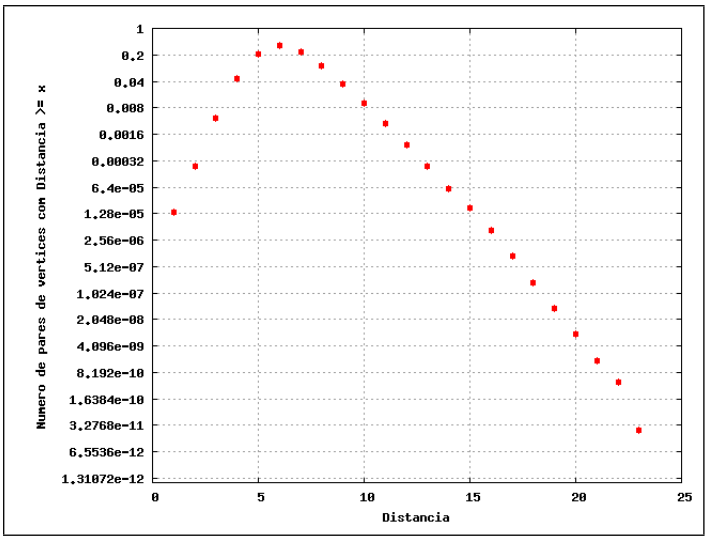


Figura 3.11: Distribuição da distância entre pares de vértices.

$$r = \max_{v_1, v_2 \in V} d(v_1, v_2) \tag{3.6}$$

3.2.5 Peso da Aresta

No capítulo 2 introduziu-se o método de Newman (ver equação 2.1) para calcular o peso das arestas de um grafo da rede de colaboração. O capítulo 2 apresenta as motivações para se utilizar esta métrica. Nesta subseção apresenta-se mais uma observação de como a normalização pelo número de colaboradores é importante para os pesos das arestas desta rede. O menor peso de aresta existente no grafo é $\frac{1}{113}$, correspondendo à um artigo com 114 autores. Este artigo forma 6441 arestas e contribui com peso 57.0 no peso total do grafo. Caso fosse utilizada a métrica simples abordada no capítulo 2, este único artigo contribuiria com 6441.0 de peso. Com o método de Newman, cada artigo contribui com a metade do número de autores para o peso total da rede, que neste caso é 57.

A média dos pesos das aresta é 0,63 e é quase 72 vezes maior do que a aresta de menor peso, então para analisar como o peso se distribui em todo o grafo, é apresentada a distribuição empírica dos pesos das arestas na figura 3.12. Nela, observa-se que o peso varia de 0,0088 a 267,77. Analisando a base DBLP, constatou-se que a aresta de maior peso é formada por dois autores que publicaram 336 artigos juntos, sendo que 224 tiveram apenas os dois como co-autores, logo, seu peso seria no mínimo 224,0. Os outros 112 artigos foram escritos com mais co-autores, correspondendo aos 43,8 de peso restante.

Outras observações importantes acerca do gráfico:

- O retângulo amarelado superior representa pessoas que colaboraram poucas vezes com muitos co-autores;

- O retângulo amarelado inferior representa pessoas que colaboraram muitas vezes com poucos co-autores;
- Existe descontinuidade em alguns pontos, como em $x = 1.0, 0.5, 0.333, 0.25, 0.2, 0.166$, pois a maioria dos autores tem artigos com 2, 3, 4, 5, 6 ou 7 autores, o que acaba aumentando a frequência das arestas com esse peso.

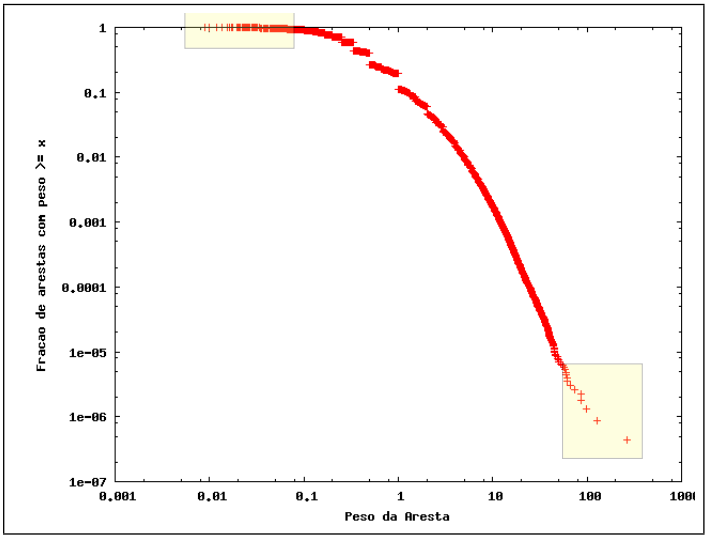


Figura 3.12: Distribuição dos pesos das arestas da rede de colaboração mundial.

3.2.6 Peso do Vértice

Assim como o peso da aresta, o peso do vértice também foi definido no capítulo 2 e é calculado através da equação 2.2. Vale lembrar que o peso do vértice corresponde ao número de publicações que o autor tem em colaboração com pelo menos um co-autor, que pode ser menor do que o número total de publicações. O peso médio de um vértice desta rede é 3,9, logo, em média, um autor possui apenas 3,9 publicações em colaboração com outros autores.

A figura 3.13 mostra a distribuição do pesos dos vértices. O menor peso de vértice é 0, porém o gráfico não mostra este ponto por estar em escala log-log, no entanto ele está contabilizado na distribuição exibida. No gráfico o menor peso de vértice é 1 e o maior é 529. Outro dado interessante visto a partir do gráfico, é que apenas 10% dos autores colaboraram em mais de 8 artigos. Como a média do peso é metade deste valor, é de se esperar que poucos autores tenham mais que o dobro da média de artigos publicados com outros co-autores. Logo, pode-se concluir que a grande maioria das pessoas colaboram muito pouco e que poucas pessoas colaboram muito.

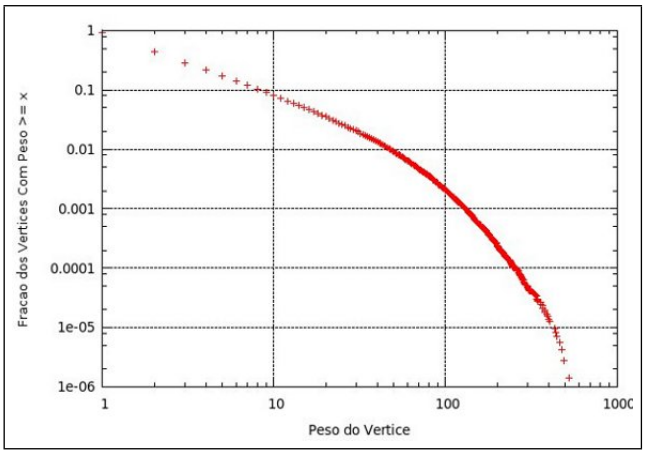


Figura 3.13: Distribuição dos pesos dos vértices da rede de colaboração mundial.

3.2.7 Idades das Publicações

A idade de uma publicação corresponde à quantidade de anos que uma publicação possui, contando desde o ano em que foi publicada até o ano da base de dados correspondente. A versão da DBLP utilizada neste trabalho é de junho de 2009, logo todos as publicações do ano de 2009 têm idade 0, as publicações do ano de 2008 tem idade 1, as de 2007 tem idade 2 e as de um ano n qualquer tem idade $2009 - n$.

Através desta métrica, encontrou-se a média e a distribuição das idades das publicações. A média é de 8,26 anos e a distribuição é apresentada na figura 3.14. As idades variam de 0 a 73 anos e existem muitas publicações com idade baixa e várias com idade alta. Por exemplo, 68% tem menos de 10 anos de idade. Isto mostra o grande crescimento recente da base de dados, elaborado pelo crescimento de publicação na área de computação ou poderia ser apenas a falta de publicações antigas cadastradas.

3.2.8 Coeficiente de Gini

A curva de Lorenz e o coeficiente de Gini, apresentados no capítulo 2 como formas de determinar o nível de desigualdade na distribuição de renda de um país, são utilizados nesta dissertação para calcular o nível de desigualdade na distribuição do peso dos vértices e do peso das arestas da rede de colaboração.

A figura 3.15 ilustra a curva de Lorenz, onde o eixo x representa a porcentagem acumulada dos vértices e o eixo y representa a porcentagem acumulada do peso dos vértices. Observando o ponto (0,8;0,27), nota-se que 80% dos indivíduos com menor peso detém apenas 27% do peso total dos vértices, ou seja, 27% de todas as publicações com ao menos um co-autor. E somente 20% dos indivíduos com maior número de publicações em colaboração possuem 73% de todas as publicações em

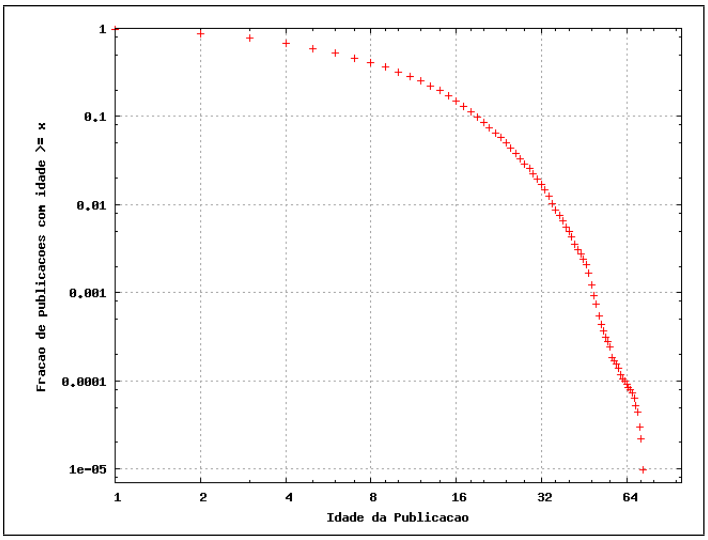


Figura 3.14: Distribuição da idade das publicações da *DBLP*.

colaboração. Desta forma, é possível ver a grande desigualdade desta distribuição, que também é representada pela curva, pois está bastante distante da diagonal. Para expressar essa desigualdade em apenas um número, utiliza-se o coeficiente de Gini. Neste caso, o coeficiente de Gini corresponde à 0,66.

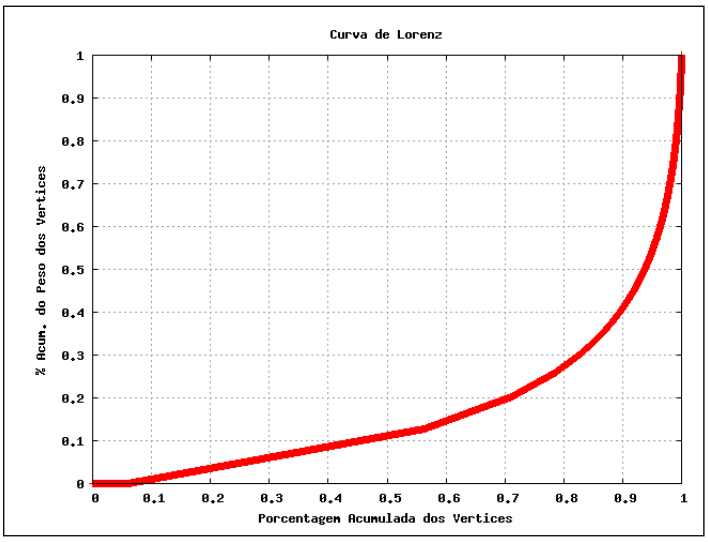


Figura 3.15: Desigualdade da distribuição do peso do vértices da rede mundial.

Já a figura 3.16 ilustra a curva de Lorenz, onde o eixo x representa a porcentagem acumulada das arestas, ou relacionamentos, e o eixo y representa a porcentagem acumulada das intensidades dos relacionamentos. Observando o ponto (0,8;0,39), nota-se que 80% das arestas de menor peso detém 39% da intensidade de relacionamento, e apenas as 20% de maior peso possuem o grande montante restante de 61% de toda intensidade de relacionamentos. A curva da figura 3.16 está mais próxima da diagonal do que a curva da figura 3.15, logo é menos desigual e possui um coeficiente de Gini de 0,55, bem menor que o anterior. Em todo o caso, fica evidente a

desigualdade de peso entre as arestas da rede de colaboração.

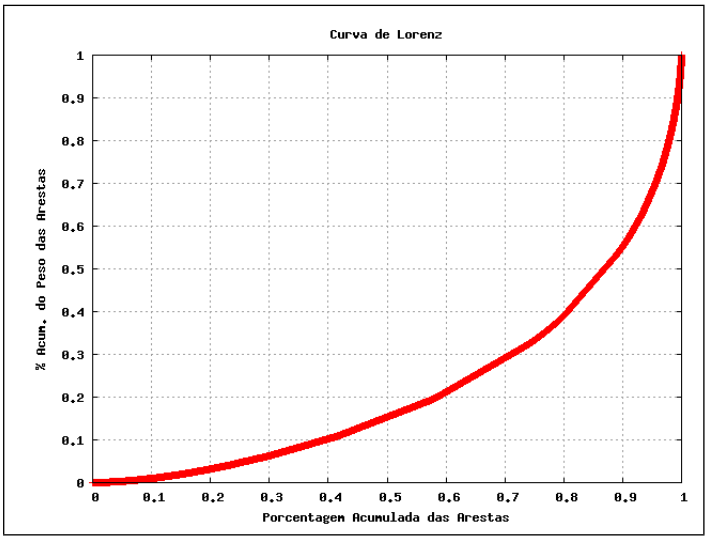


Figura 3.16: Desigualdade da distribuição do peso das arestas da rede mundial.

3.2.9 Número de Co-autores e Número de Publicações

Com a motivação de descobrir se existe uma correlação entre o número de colaboradores e o número de publicações de um pesquisador, plotou-se o grau *versus* número de publicações ilustrado na figura 3.17. Para ser plotada, o grau e o número de publicações foi dividido na forma de histograma em 20 intervalos de tamanhos exponencialmente maiores. O eixo x representa o grau de um vértice, o eixo y representa o número de publicações de um vértice e o eixo z é a quantidade de vértices com grau x e y publicações. Observando o gráfico, nota-se que ocorre uma concentração de uma grande quantidade de nós na faixa diagonal e que, em geral, há uma tendência a quanto maior o número de co-autores, maior o número de publicações. Entretanto existem algumas exceções de vértices com alto grau e baixo número de publicações, e de vértices com grande quantidade de publicações e baixo grau, por exemplo, existem dois vértices com apenas 1 colaborador e 214 publicações. Existem, também, quarenta vértices com 1 publicação e 107 colaboradores.

3.3 Análise da Rede Brasileira

Nesta seção, são apresentadas as propriedades topológicas da rede de colaboração quando considerados apenas o conjunto de pessoas que atuam no Brasil, ou seja, a rede de colaboração brasileira. Considera-se assim o subgrafo induzido obtido a partir da rede de colaboração mundial analisada na seção 3.2. Ele contém somente 2.729 vértices e 6.953 arestas entre eles, todas os demais vértices e arestas foram desconsiderados.

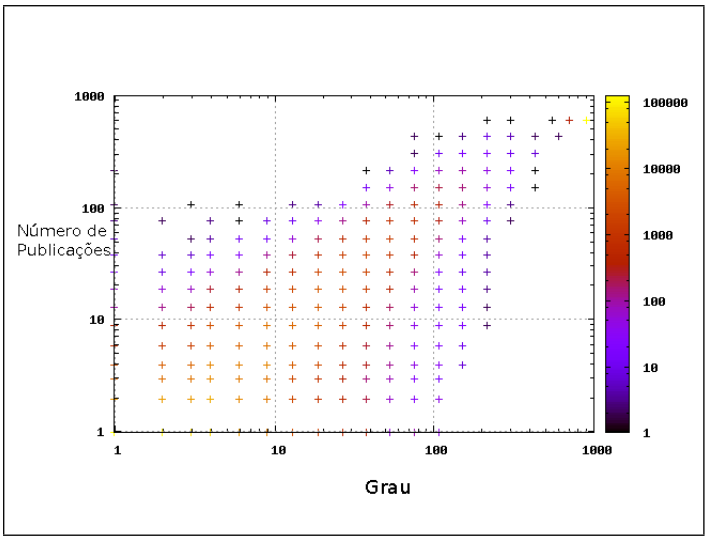


Figura 3.17: Número de co-autores *versus* número de publicações de cada pessoa da rede mundial.

As mesmas métricas serão utilizadas para caracterizar esta rede. Ao final do capítulo, é apresentada uma tabela comparativa apresentando o resumo das métricas das duas redes de colaboração.

3.3.1 Grau

A rede de colaboração brasileira possui um grau médio menor do que a rede mundial. Enquanto na rede mundial, um autor tem em média 6,3 colaboradores, um pesquisador que atua no Brasil tem em média 5,1 colaboradores que atuam no Brasil, variando de 0 a 101. Logo, os indivíduos que atuam no Brasil têm em média menos colaboradores do que os atuantes em outros países. A distribuição empírica do grau deste grafo é apresentada na figura 3.18 e mostra-se similar à distribuição do grau da rede mundial. Há um pequeno número de autores com muitas colaborações e uma grande quantidade de pessoas com poucos co-autores, por exemplo, no gráfico observa-se que apenas 11% dos vértices têm grau maior ou igual a 10 e que 80% dos vértices têm grau menor ou igual a 8.

Através da distribuição do grau, notou-se também uma característica não esperada: 9% dos indivíduos que atuam no Brasil não têm colaboração, eles correspondem a 243 brasileiros que publicaram sem colaboradores que atuam no Brasil e por isto não possuem arestas na rede de colaboração brasileira.

3.3.2 Componentes Conexas

Através da figura 3.19 pode-se constatar que existe uma componente gigante e várias outras muito pequenas. Na componente gigante estão quase todos os vértices

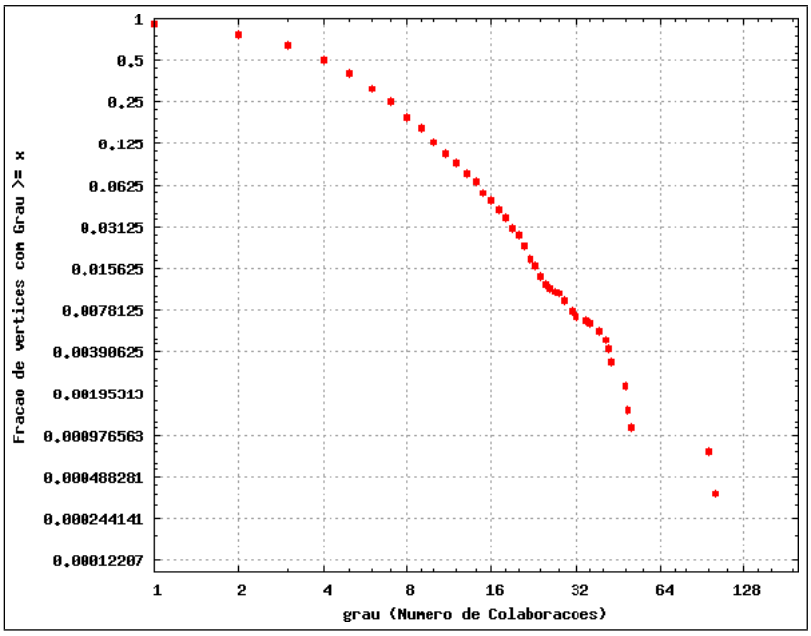


Figura 3.18: Distribuição do grau na rede de colaboração brasileira.

do grafo, 2.338 vértices, constituindo 85,7% da rede de colaboração brasileira, e a segunda maior componente é pequena, com apenas 13 vértices.

As componentes conexas da rede brasileira têm, em média, o tamanho igual a 9,2 e no total são 297 componentes. Como é um grafo com muitos vértices (2.729), essa média indica que existem muitas componentes com poucos vértices. A rede brasileira tem uma distribuição do tamanho das componentes conexas semelhante à rede mundial, pois a rede mundial também possui uma componente gigante relativamente próxima, 79,8%, e uma média de 9,3 no tamanho das componentes conexas. Para esclarecer, o gráfico da distribuição do tamanho das componentes conexas é apresentado na figura 3.19.

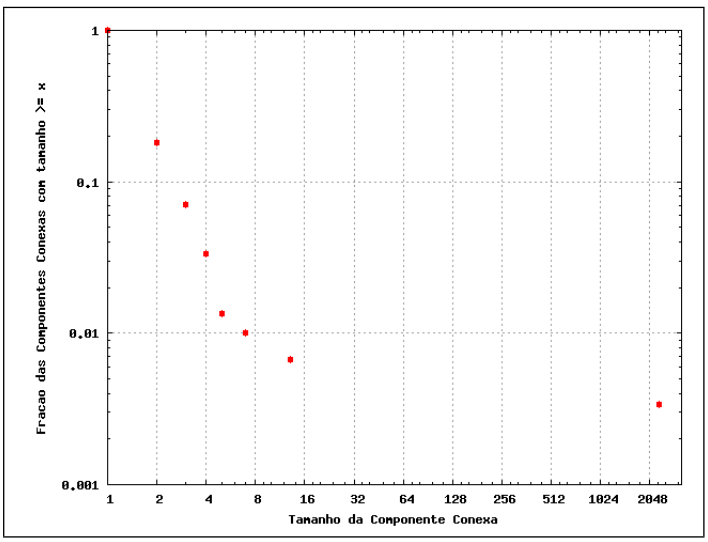


Figura 3.19: Distribuição do tamanho das componentes conexas.

3.3.3 Coeficiente de Clusterização

O coeficiente de clusterização da rede de colaboração brasileira é 0.48, 19% menor que o da rede mundial que é 0,59. Logo, a chance de existir uma colaboração entre duas pessoas que possuem um colaborador em comum é menor do que na rede mundial. Entretanto, este valor ainda é relativamente alto, pois existe quase 50% de chance dessa colaboração existir.

A distribuição do coeficiente de clusterização está na figura 3.20, a partir dela, observa-se que 33% dos vértices tem coeficiente de clusterização maior que 0,7 e que os 26% dos vértices com coeficiente de clusterização mais baixo têm um valor menor ou igual a 0,1. Portanto, podemos concluir que a maioria dos vértices da rede estão envolvidos em triângulos, ou seja, altamente conectados no interior do grafo.

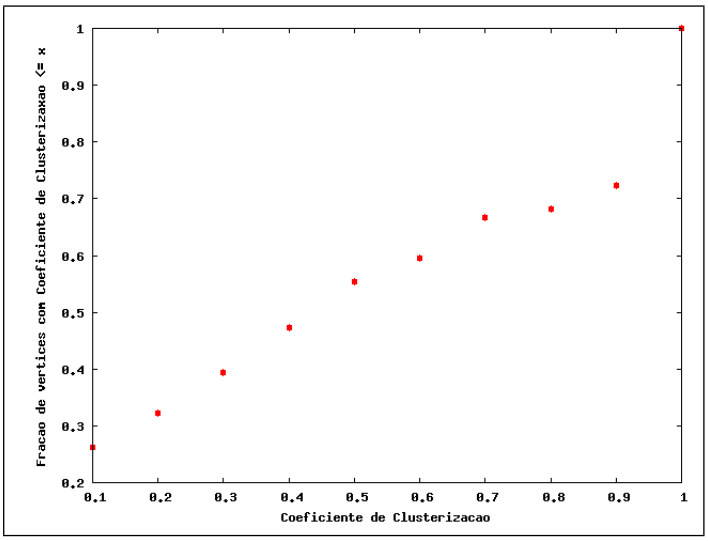


Figura 3.20: Distribuição do coeficiente de clusterização da rede de colaboração brasileira.

3.3.4 Distância

Assim como a rede mundial, a rede de colaboração brasileira também possui o efeito *mundo pequeno*, pois possui uma distância média pequena entre pares de vértices, de apenas 5,6, e uma alta clusterização de 0,48.

A distribuição da distância está na figura 3.21 e através dela fica evidente sua média, observando que mais de 51% dos pares de vértices tem distância 5 e 6. As distâncias se concentram em caminhos curtos, pares de nós com distâncias 4 e 7 somam 31%. Somando-os aos 51% anteriores, totaliza em 82% do grafo sendo formado por pares de vértices com distâncias 4, 5, 6 e 7. Além disso, ela mostra que as distâncias bem maiores existem, mas com baixas frequências. O diâmetro da rede, ou seja, o maior caminho mínimo entre dois vértices da rede de colaboração brasileira é 15, portanto, menor que o da rede mundial que é de 23. Logo, o maior

caminho mínimo entre dois vértices da rede de colaboração mundial não é entre dois pesquisadores que atuam no Brasil.

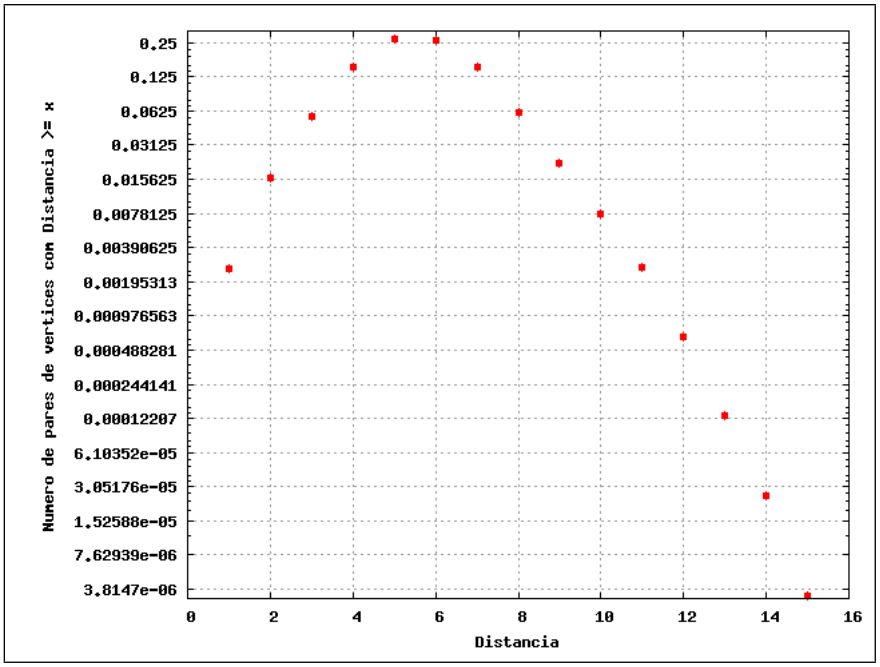


Figura 3.21: Distribuição das distâncias na rede de colaboração brasileira.

3.3.5 Peso das Arestas

A rede de colaboração brasileira possui o peso médio das arestas igual a 1,09. Uma única publicação não contribui mais de 1,0 no peso de uma aresta, logo, pode-se concluir que dois autores que colaboraram tem, em média, mais de uma publicação juntos. Observando a figura 3.22 percebe-se que a distribuição do peso das arestas possui uma cauda pesada, ocorrendo valores muito distantes do peso médio, variando de 0,03 a 86,6. A aresta de maior peso (86.6) deste grafo é única e três vezes menor que a da rede de colaboração mundial. Esta aresta é formada por dois autores que publicaram 100 artigos em conjunto, dos quais 80 foram escritos em colaboração apenas dos dois. A aresta de menor peso também é única, e é formada por dois pesquisadores que publicaram somente um artigo com 29 co-autores.

Outras observações importantes acerca da distribuição dos pesos das arestas:

- Os primeiros pontos representam pessoas que colaboraram poucas vezes com muitos co-autores;
- Os últimos pontos representam pessoas que colaboraram muitas vezes com poucos co-autores;
- Assim como no gráfico da distribuição dos pesos das arestas da rede de colaboração mundial, a descontinuidade continua neste gráfico, pois a maioria

dos autores tem artigos com 2, 3, 4 ou 5 autores, o que acaba aumentando a frequência das arestas com peso 1,0, 0,5, 0,333 e 0,25.

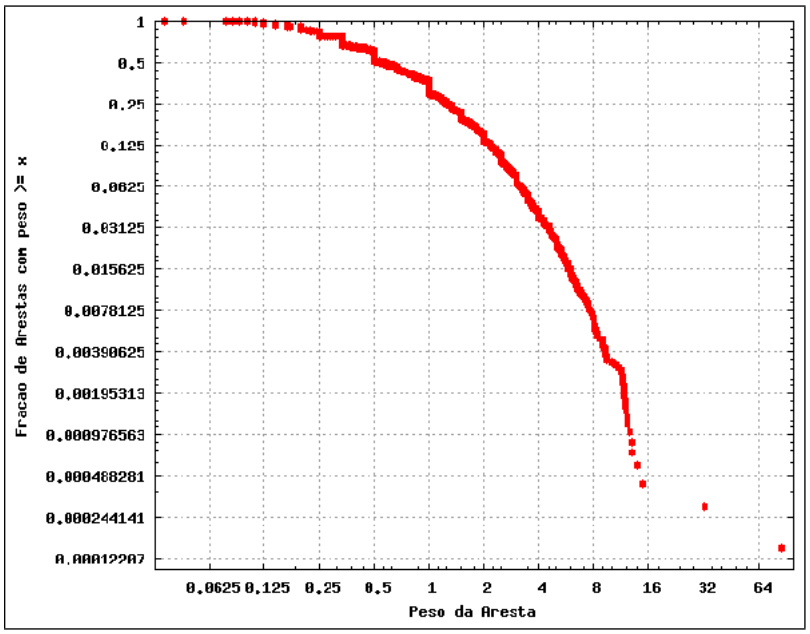


Figura 3.22: Distribuição dos pesos das arestas na rede de colaboração brasileira.

3.3.6 Peso dos Vértices

O peso médio de um vértice desta rede é 5,6, logo, em média, um autor possui 5,6 publicações em colaboração com outros autores, um valor médio maior que na rede de colaboração mundial.

Observando a figura 3.23 percebe-se que a distribuição do peso dos vértices possui uma cauda pesada, ocorrendo valores muito distantes do peso médio, variando de 0 a 123. Na rede brasileira, vértices com peso igual a zero correspondem às pessoas que possuem todas as suas publicações sem nenhum co-autor que atua no Brasil. Outro dado interessante visto a partir do gráfico, é que apenas 19% dos pesquisadores publicaram mais de 8 artigos com ao menos um co-autor, enquanto que na rede mundial, apenas 10% dos autores colaboraram em mais de 8 artigos. Logo, os pesquisadores que atuam no Brasil tendem a publicar mais artigos em colaboração que os pesquisadores de outros países.

3.3.7 Idades das Publicações

Utilizando a mesma métrica introduzida na subseção 3.2.7 para calcular a idade das publicações dos autores da rede mundial, calculou-se a média e a distribuição da idade das publicações dos autores que atuam no Brasil. A média é de 5,46 anos, bem menor que na rede mundial, que é de 8,26. Já a distribuição é apresentada

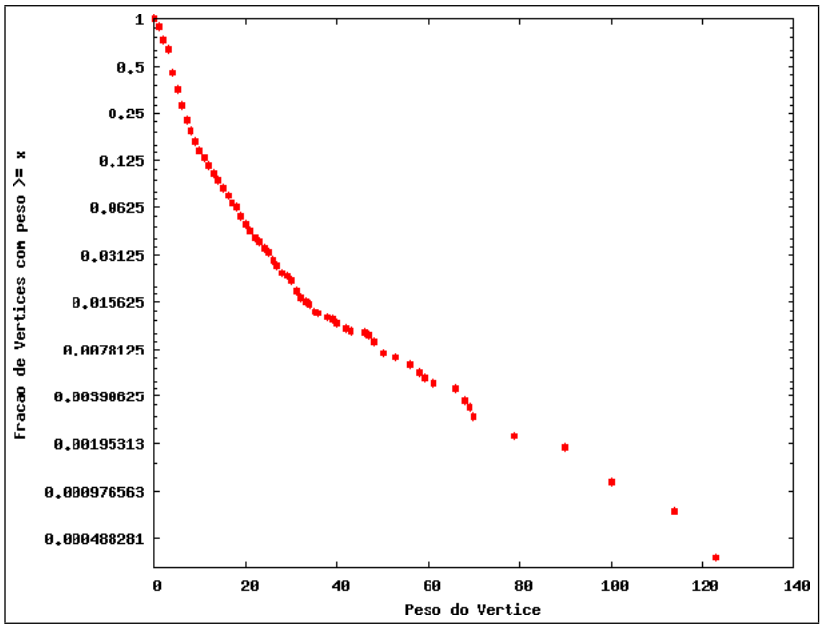


Figura 3.23: Distribuição dos pesos dos vértices na rede de colaboração brasileira.

na figura 3.24. A partir da distribuição, percebe-se que a maior idade de uma publicação na rede brasileira é bem menor do que na rede mundial, pois enquanto a publicação mais antiga da rede mundial tem 73 anos, a publicação brasileira mais antiga presente nesta base de dados tem apenas 38 anos. Também percebe-se uma característica em comum com as idades das publicações da rede mundial, pois na rede brasileira existem muitas publicações com idade baixa e poucas com idade alta. Por exemplo, 85% tem menos de 10 anos de idade. Entretanto, apenas 68% das publicações da rede mundial tem idade menor que 10 anos. Logo, a rede brasileira é mais nova, pois as publicações são mais recentes.

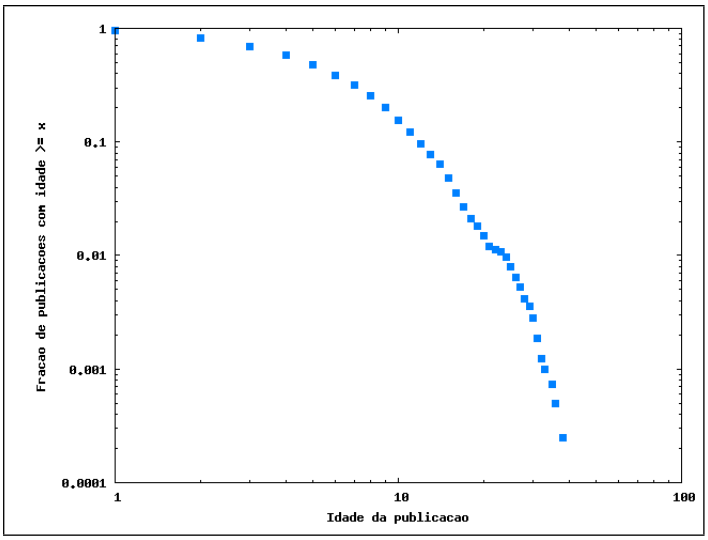


Figura 3.24: Distribuição da idade das publicações dos autores que atuam no Brasil.

3.3.8 Coeficiente de Gini

A figura 3.25 ilustra a curva de Lorenz, onde o eixo x representa a porcentagem acumulada dos vértices e o eixo y representa a porcentagem acumulada do peso dos vértices. Observando o ponto (0,8;0,38), nota-se que 80% dos indivíduos com menor peso detém apenas 38% do peso total dos vértices, ou seja, 38% de todas as publicações com ao menos um co-autor. E somente 20% dos indivíduos com maior número de publicações em colaboração possuem 62% de todas as publicações em colaboração. Desta forma, é possível ver a grande desigualdade desta distribuição, que também é representada pela curva, pois está bastante distante da diagonal. Para expressar essa desigualdade em apenas um número, utiliza-se o coeficiente de Gini. Neste caso, o coeficiente de Gini corresponde à 0,58. Logo, a distribuição do peso dos vértices da rede brasileira é menos desigual do que a distribuição do peso dos vértices na rede mundial, pois esta possui o coeficiente de Gini igual a 0,66.

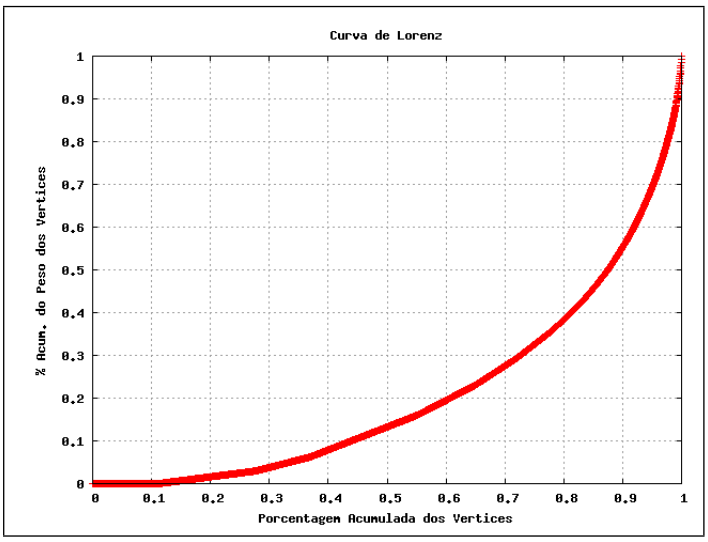


Figura 3.25: Desigualdade da distribuição do peso dos vértices da rede brasileira.

Já a figura 3.26 ilustra a curva de Lorenz, onde o eixo x representa a porcentagem acumulada das arestas, ou relacionamentos, e o eixo y representa a porcentagem acumulada das intensidades dos relacionamentos. Observando o ponto (0,8;0,42), nota-se que 80% das arestas de menor peso detém 39% da intensidade de relacionamento, e apenas as 20% de maior peso possuem o grande montante restante de 58% de toda intensidade de relacionamentos. A desigualdade da distribuição representada pela curva da figura 3.26 está semelhante à desigualdade da distribuição do peso das arestas da rede mundial representada pela curva da figura 3.16, logo seus coeficientes de Gini são próximos. O coeficiente de Gini na rede brasileira é 0,54, enquanto que o coeficiente de Gini na rede mundial é 0,55.

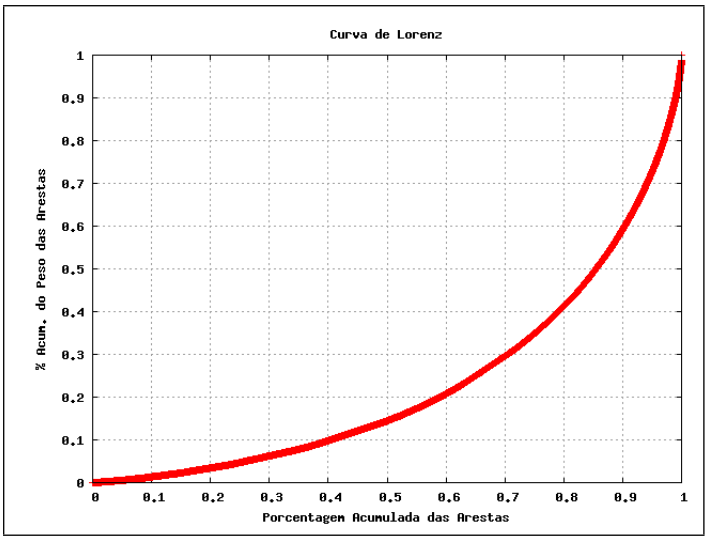


Figura 3.26: Desigualdade da distribuição do peso das arestas da rede brasileira.

3.4 Resumo das Métricas

A Tabela 3.1 apresenta os resultados numéricos que foram discutidos em detalhes nas seções anteriores. A única métrica que não está contida em seções anteriores é o número médio de publicações por vértice. A métrica foi calculada através da razão entre o número total de publicações e o número de vértices.

Ao observar as métricas peso médio do vértice e grau médio da rede mundial e da rede brasileira, constata-se que o peso médio do vértice é muito maior na rede brasileira, mas o grau é menor. Logo, os pesquisadores que atuam no Brasil têm, em média, mais publicações em colaboração do que o restante do mundo, porém possuem menos colaboradores, uma vez que o grau médio é menor.

Tabela 3.1: Resumo das Métricas

	Rede Mundial	**Rede Brasileira**
Número de vértices	722.392	2.729
Número de arestas	2.272.540	6.953
Número de publicações	1.230.213	13.314
Núm. médio de publicações/vértice	1,7	4,9
Grau médio	6,3	5,1
Tam. da comp. gigante	576.309	2.338
Tam. da comp. gigante (em %)	79,8%	85,7%
Tam. da 2º maior comp. gigante	42	13
Número de componentes conexas	77.493	297
Tam. médio das comp. conexas	9,3	9,2
Coeficiente de clusterização	0,59	0,48
Distância média	6,3	5,6
Diâmetro	23	15
Peso médio das arestas	0,63	1,09
Peso médio do vértice	3,9	5,6
Idade média das publicações	8,3	5,5
Gini (distrib. do peso dos vértices)	0,66	0,58
Gini (distrib. do peso das arestas)	0,55	0,54

Capítulo 4

Métrica para Ranqueamento Baseada em Intensidade de Relacionamento

Este capítulo apresenta a principal contribuição desta dissertação. Será introduzida uma métrica para ranqueamento baseada na intensidade de relacionamento e utilizá-la para avaliar programas de pós-graduação do Brasil na área de Ciência da Computação, fazendo uma comparação de seus resultados com a avaliação subjetiva feita pela CAPES e com várias métricas definidas nos capítulos anteriores. A métrica proposta será utilizada para avaliar os pesquisadores que atuam no Brasil, fazendo uma comparação de seus resultados com a avaliação subjetiva feita pelo CNPq e outras métricas definidas nos capítulos anteriores.

A intensidade dos relacionamentos está relacionada com a importância do vértice ou do conjunto de vértices, pois representa a influência deste sobre o grafo. Por exemplo, a capacidade de comunicação deste subconjunto com os outros vértices do grafo é proporcional à intensidade dos relacionamentos.

4.1 Pesos e Cortes

A métrica proposta utiliza a ideia de corte e de peso do corte em um grafo. Em um grafo qualquer $G = (V, E)$, o corte associado a um conjunto X de vértices é o conjunto de todas as arestas que têm uma ponta em X e a outra em $V - X$, onde V é o conjunto de todos os vértices pertencentes ao grafo G [37]. O peso do corte é dado pela soma dos pesos das arestas que definem o corte. Ou seja, dado um conjunto X de vértices, corresponde à soma dos pesos de todas as arestas que têm uma ponta em X e a outra em $V - X$. Logo, os pesos das arestas formadas entre os vértices do conjunto X não contribuem no peso do corte. Já o peso do vértice

no corte é dado pela soma dos pesos das arestas que definem o corte e incidem no vértice. Para utilizar esta métrica precisamos de um grafo G e um conjunto X de vértices, pois a partir destes dois dados, é possível obter o peso do corte. A Figura 4.1 será utilizadas para exemplificar como medir o peso do corte e o peso do vértice no corte.

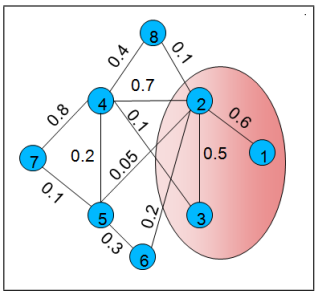


Figura 4.1: Exemplo de cálculo do peso do corte.

A figura 4.1 ilustra um grafo G com sete vértices enumerados de 1 a 7. O conjunto X de vértices, dado como entrada para o cálculo do peso dos vértices no corte, está identificado por uma elipse. Ele corresponde aos vértices 1, 2 e 3. O cálculo do peso dos vértices no corte pode ser feito a partir dos vértices de dentro do conjunto ou dos vértices de fora do conjunto. Calculando o peso dos vértices no corte a partir dos vértices de dentro do conjunto, obtemos que o peso dos vértices 1, 2 e 3 no corte é igual a 0, 1,05 e 0,1 respectivamente. Ao calcularmos este peso a partir dos vértices de fora do conjunto X, obtemos que o peso dos vértices 4, 5, 6, 7 e 8 no corte é igual a 0,8, 0,05, 0,2, 0 e 0,1 respectivamente. O peso do corte no grafo corresponde à soma destes pesos (de dentro ou de fora) resultando em 1,15.

Com estes dados obtidos, pode-se fazer a distribuição empírica da contribuição de peso dos vértices no corte de dentro e de fora do conjunto, além de gerar a média e o desvio-padrão. É importante ressaltar que a soma dos pesos dos vértices no corte de dentro é igual à soma dos pesos dos vértices de fora, porém suas médias são diferentes. Por exemplo, na figura 4.1 existem 3 vértices no conjunto, logo a média do peso dos vértices de dentro do conjunto é $\frac{1,15}{3}$, enquanto que a média do peso dos vértices de fora do conjunto é $\frac{1,15}{4}$, pois apenas quatro vértices de possuem arestas para vértices de dentro do conjunto.

Também é importante salientar que diferentes vértices contribuem de forma diferente para o peso do corte. No exemplo acima, o vértice 2 contribui com peso 1,05 para o corte que tem peso total de 1,15. Ou seja, 91% do peso do corte vem do vértice 2. Essa desigualdade na distribuição empírica do peso dos vértices no corte pode ser medida através do coeficiente de Gini, introduzido na seção 2.1.

Como a intensidade dos relacionamentos está relacionada com a importância do vértice, podemos ranquear os vértices internos ao conjunto de acordo com seu peso

no corte. Assim, consideraremos que o vértice mais importante é aquele que tem maior peso do vértice no corte. Logo, no exemplo anterior, o vértice mais importante do conjunto é o vértice 2, por ter o maior peso. O segundo mais importante é o vértice 3 por ter o segundo maior peso.

4.2 Ranqueamento dos Programas de Pós-graduação do Brasil

Nesta seção busca-se a métrica objetiva que faz o melhor ranqueamento dos programas de pós-graduação da área da ciência da computação do Brasil, comparando os resultados obtidos com a avaliação dos programas realizada pela CAPES. A motivação desta comparação é verificar se a métrica objetiva baseada em peso do corte, proposta na seção 4.1 faz um melhor ranqueamento do que as outras métricas objetivas.

4.2.1 Caracterização dos conjuntos

Para fazer o ranqueamento dos programas de pós-graduação a partir de métricas objetivas aplicadas à rede de colaboração, foi necessário, primeiramente, identificar os programas na rede. Cada programa de pós-graduação é formado por um conjunto de professores, e cada professor corresponde a um vértice no grafo de colaboração.

Os professores vinculados aos programas de pós-graduação foram obtidos do Sistema de Informação CAPES e CNPq, disponível em [38], identificados na DBLP e separados em conjuntos. Ao todo foram avaliados vinte e um programas de pós-graduação na área da Ciência da Computação.

As métricas foram aplicadas à rede de colaboração mundial utilizando um conjunto de cada vez. Para constatar a eficácia das métricas, é necessário comparar os resultados com uma avaliação existente. A avaliação da pós-graduação feita pela CAPES foi utilizada como base para identificar as métricas que melhor ranqueiam os programas de pós-graduação de acordo com seu nível de qualidade.

O Sistema de Avaliação da Pós-graduação da CAPES existe desde 1976 e desde então vem cumprindo papel de fundamental importância para o desenvolvimento da pós-graduação e da pesquisa científica e tecnológica no Brasil [39]. A Avaliação dos programas de pós-graduação é feita através de um acompanhamento anual e da avaliação trienal do desempenho de todos os programas [39]. São atribuídas notas de 1 a 7, nas quais os programas com nota 6 ou 7 são os programas de excelência na área.

Foram escolhidos 21 programas de pós-graduação aleatoriamente. A distribuição das notas dos programas analisados neste trabalho de acordo com a avaliação da

CAPES é:

- dois programas com nota 7;
- três programas com nota 6;
- quatro programas com nota 5;
- oito programas com nota 4;
- quatro programas com nota 3.

Os nomes dos programas foram mantidos em anonimato, mas os dados correspondem a programas reais. Os programas de mesma nota são denominados pela sua respectiva nota CAPES, acompanhada por letras do alfabeto para serem identificados, sem que isto tenha relacão alguma com o ranqueamento.

4.2.2 Avaliação dos Conjuntos

Para fazer a avaliação dos conjuntos, várias métricas foram consideradas. Os resultados são apresentados nas tabelas 4.1 e 4.2. Abaixo segue a explicação das métricas apresentadas em cada uma das colunas das tabelas:

- Nota CAPES: nota atribuída subjetivamente pela CAPES para classificar os programas em níveis de qualidade, explicada em detalhes na seção 4.2.1;
- Número de autores (# de autores): quantidade de pesquisadores (vértices) que fazem parte do programa e que estão cadastrados na DBLP;
- Número de arestas do subgrafo induzido (# de arestas do subgrafo induzido): número de arestas existentes apenas entre os vértices do programa considerando a rede de colaboração estudada;
- Grau médio do subgrafo induzido: duas vezes o número de arestas do subgrafo induzido dividido pelo número de vértices do conjunto;
- Número de componentes conexas do subgrafo induzido (# de CC do subgrafo induzido): quantidade de componentes conexas do subgrafo contendo apenas os vértices do conjunto;
- Tamanho da maior componente conexa do subgrafo induzido (Tamanho da maior CC do subgrafo): quantidade de vértices na maior componente conexa do sugbrafo induzido;

Tabela 4.1: Métricas aplicadas aos programas de pós-graduação - parte 1.

Programa	Nota CAPES	# de autores (nós)	# de arestas do Subgrafo Induzido	Grau médio do subgrafo Induzido	# CC do subgrafo Induzido	Tamanho Da maior CC do subgrafo	Tam. Da maior CC normalizado pelo #vertic.	Grau dos vértices (Soma)	Grau médio dos vértices	#Pub	# Médio de Pub por vértice	Peso do Programa	Peso médio do vértice no grupo	Peso do corte	Número de arestas do corte	Média (corte das arestas)	Desvio-padrão (corte das arestas)	Peso médio do corte (dentro)	Desvio-padrão (corte dos vértices – dentro)	# vértices do corte (Fora)	Peso médio do corte (fora)	Desvio-padrão (corte dos vértices – fora)	Peso do corte (exterior sem os brasileiros)	Peso médio do corte (sem os br. - dentro)	Std (corte dos vértices sem br. - dentro)	# vértices do corte (fora e sem os brasileiros)	Peso médio do corte (sem os br. - fora)	Std (corte dos vértices sem br. - fora)	Coef. de Gini (dentro)	Coef. de Gini (Fora)
7A	7	40	54	2.7	7	33	0.83	1202	30.05	900	22.5	1061	26.53	880.63	1094	0.8	3.83	22.02	22.08	826	1.07	4.28	419.23	10.48	11.43	560	0.75	2.57	0.51	0.55
7B	7	24	48	4	2	22	0.92	977	40.71	855	35.63	993	41.38	828.16	881	0.94	4.38	34.51	32.28	454	1.22	4.98	368.02	15.33	15.03	454	0.81	2.7	0.4	0.59
6A	6	28	47	3.36	5	24	0.86	1276	45.57	773	27.61	1057	37.75	861.34	1182	0.73	3.72	30.76	27	745	1.16	4.68	426.25	15.22	12.9	532	0.8	2.58	0.46	0.56
6B	6	46	69	3	8	39	0.85	1207	26.24	886	19.26	1027	22.33	873.75	1069	0.82	4.1	18.99	19.11	809	1.08	4.58	455.43	9.9	7.97	625	0.73	2.59	0.48	0.54
6C	6	54	104	3.85	7	49	0.91	1961	36.31	1291	23.91	1666	30.85	1329.5	1753	0.76	4.69	24.62	20	1168	1.14	5.66	680.56	12.6	9.87	887	0.77	3.02	0.43	0.53
5A	5	48	48	2	17	38	0.79	1005	20.94	792	16.5	919	19.15	777.12	909	0.85	3.93	16.19	17.57	708	1.1	4.27	401.71	9.16	10.22	531	0.83	2.73	0.49	0.5
5B	5	23	16	1.39	10	12	0.52	434	18.87	299	13	353	15.35	294.31	402	0.73	2.16	12.8	11.43	309	0.95	2.77	161.43	7.02	7.16	225	0.72	1.55	0.42	0.49
5C	5	43	49	2.28	13	18	0.42	654	15.21	541	12.58	602	14	476.93	556	0.86	3.11	11.09	13.04	432	1.1	3.64	344.61	8.01	9.87	365	0.94	2.83	0.53	0.5
5D	5	46	55	2.39	11	35	0.76	1011	21.98	687	14.93	915	19.89	723.61	901	0.8	3.89	15.73	21.82	638	1.13	4.52	344.95	7.5	9.93	462	0.75	2.38	0.58	0.54

Tabela 4.2: Métricas aplicadas aos programas de pós-graduação - parte 2.

Programa	Nota CAPES	# de autores (nós)	# de arestas do Subgrafo Induzido	Grau médio do subgrafo induzido	# CC do subgrafo Induzido	Tamanho da maior CC do subgrafo	Tam. Da maior CC normalizado pelo #vertic.	Grau dos vértices (Soma)	Grau médio dos vértices	#Pub	# Médio de Pub por vértice	Peso do Programa	Peso médio do vértice no grupo	Peso do corte	Número de arestas do corte	Média (corte das arestas)	Desvio-padrão (corte das arestas)	Peso médio do corte (dentro)	Desvio-padrão (corte dos vértices – dentro)	# vértices do corte (Fora)	Peso médio do corte (fora)	Desvio-padrão (corte dos vértices – fora)	Peso do corte (exterior sem os brasileiros)	Peso médio do corte (sem os br. - dentro)	Std (corte dos vértices sem br. - dentro)	# vértices do corte (fora e sem os brasileiros)	Peso médio do corte (sem os br. - fora)	Std (corte dos vértices sem br. - fora)	Coef. de Gini (dentro)	Coef. de Gini (Fora)
4A	4	17	7	0.82	11	3	0.18	212	12.47	132	7.76	154	9.06	132.67	198	0.67	1.58	7.8	6.44	175	0.76	1.87	71.53	4.21	5.44	126	0.57	0.73	0.42	0.52
4B	4	21	22	2.1	5	17	0.81	557	26.52	274	13.05	351	16.71	281.97	513	0.55	1.49	13.35	12.06	380	0.74	1.76	185.06	8.81	7.93	303	0.61	1.18	0.45	0.5
4C	4	14	2	0.29	12	2	0.14	136	9.71	108	7.71	109	7.79	108.25	132	0.82	1.52	7.73	8.89	119	0.91	1.68	46.45	3.32	4.18	72	0.65	0.52	0.51	0.53
4D	4	27	19	1.41	11	8	0.3	491	18.19	332	12.3	415	15.37	316.33	453	0.7	2.07	11.72	15.01	350	0.7	2.07	186.25	6.9	8.25	265	0.7	1.7	0.51	0.55
4E	4	19	9	0.95	10	6	0.32	345	18.16	244	12.84	284	14.95	242.83	327	0.74	1.76	12.19	11.12	262	0.93	1.98	119.43	6.29	4.64	174	0.69	0.98	0.4	0.46
4F	4	20	22	2.2	3	18	0.9	403	20.15	285	14.25	335	16.75	295.3	359	0.82	2.14	14.77	10.54	286	1.03	2.41	171.53	8.58	10.33	196	0.88	2	0.35	0.48
4G	4	18	19	2.11	8	7	0.39	270	15	244	13.56	242	13.44	188.27	232	0.81	2.09	10.46	7.73	169	1.11	1.93	128.62	7.15	5.06	125	1.03	1.7	0.41	0.44
4H	4	15	5	0.67	10	5	0.33	223	14.87	173	11.53	185	12.33	168.47	213	0.79	1.84	11.23	11.38	188	0.9	1.99	89.71	5.98	6.63	136	0.66	1.09	0.47	0.48
3A	3	12	2	0.33	10	2	0.17	148	12.33	72	6	85	7.08	77.1	144	0.54	0.26	6.42	6.3	121	0.64	0.71	65.86	5.49	5.73	113	0.58	0.41	0.45	0.43
3B	3	15	7	0.93	9	3	0.2	321	21.4	191	12.73	219	14.6	195.61	307	0.64	1.88	13.04	16.12	236	0.83	2.03	85.88	5.73	4.87	149	0.58	0.63	0.6	0.55
3C	3	10	9	1.8	4	7	0.7	195	19.5	139	13.9	165	16.5	143.93	177	0.81	1.76	14.39	9.28	145	0.99	2	116.5	11.65	8.33	123	0.95	116.5	0.36	0.5
3D	3	24	16	1.33	11	6	0.25	205	8.54	120	5	136	5.67	114.03	173	0.66	1.62	4.75	6.53	139	0.82	1.83	84.8	3.53	5.34	103	0.82	1.64	0.58	0.49

- Tamanho da maior componente conexa normalizado pelo número de vértices (Tam. Da maior CC norm. pelo #vertic.): tamanho da maior componente conexa dividido pelo número de vértices do subgrafo induzido;
- Grau dos vértices (Soma): soma dos graus de cada vértice do conjunto, considerando as arestas de toda a rede de colaboração;
- Grau médio dos vértices: soma dos graus dos vértices do conjunto dividido pelo número de vértices do conjunto;
- Número de publicações (#Pub): total de publicações de todos os vértices do conjunto;
- Número médio de publicações por vértice (# Médio de Pub por vértice): número de publicações dividido pela quantidade de vértices do conjunto;
- Peso do programa: soma dos pesos das arestas incidentes a cada um dos vértices do conjunto;
- Peso médio do vértice no grupo: peso do programa dividido pelo número de vértices do conjunto;
- Peso do corte: soma dos pesos das arestas que definem o corte quando utiliza-se o conjunto de vértices que definem o programa;
- Número de arestas do corte: quantidade de arestas que possuem uma ponta dentro do conjunto e outra ponta fora;
- Média (corte das arestas): peso médio das arestas no corte, ou seja, a soma dos pesos das arestas que tem uma ponta dentro do conjunto e outra fora dividido pela quantidade dessas arestas;
- Desvio-padrão (corte das arestas): desvio-padrão dos pesos das arestas no corte;
- Peso médio do corte (dentro): peso das arestas que conectam vértices do conjunto aos vértices de fora do mesmo dividido pela quantidade de vértices no conjunto;
- Desvio-padrão (corte dos vértices - dentro): desvio-padrão dos pesos dos vértices de dentro no corte;
- Número de vértices do corte (fora): quantidade de vértices de fora do conjunto que possuem aresta para o lado de dentro;

- Peso médio do corte (fora): peso das arestas que conectam vértices de fora do conjunto com vértices de dentro dividido pela quantidade de vértices de fora do conjunto que contribuem para o peso do corte;
- Desvio-padrão (corte dos vértices - fora): desvio-padrão dos pesos dos vértices de fora do conjunto no corte;
- Peso do corte (exterior sem os brasileiros): soma dos pesos das arestas que ligam os vértices de dentro do conjunto aos vértices de fora, com exceção das arestas que chegam até vértices que representam pesquisadores que atuam no Brasil;
- Peso médio do corte (sem os brasileiros - dentro): peso das arestas que saem dos vértices do conjunto e possuem a outra ponta em um pesquisador que não atua no Brasil dividido pela quantidade de vértices no conjunto;
- Desvio-padrão (Std - corte dos vértices sem os brasileiros - dentro): desvio-padrão dos pesos dos vértices do conjunto no corte. Esta métrica não considera os pesos das arestas com pesquisadores que atuam no Brasil;
- Número de vértices do corte (fora e sem brasileiros): quantidade de vértices de fora do conjunto que possuem aresta para o lado de dentro e que não atuam no Brasil;
- Peso médio do corte (sem os brasileiros - fora): peso das arestas que conectam pesquisadores que atuam no exterior com pesquisadores que atuam no Brasil dividido pela quantidade de pesquisadores do exterior que tem relação com o Brasil;
- Desvio-padrão (Std - corte dos vértices sem os brasileiros - fora): desvio-padrão dos pesos dos vértices de fora do conjunto no corte. Estes pesos não consideram os pesos das arestas com pesquisadores que atuam no Brasil;
- Coeficiente de Gini (dentro): coeficiente de Gini calculado a partir dos vértices de dentro do conjunto;
- Coeficiente de Gini (fora): coeficiente de Gini calculado a partir dos vértices de fora do conjunto que se relacionam com os de dentro.

Analisando as tabelas 4.1 e 4.2, observou-se que nenhuma métrica objetiva aplicada à rede de colaboração foi capaz de reproduzir o ranqueamento subjetivo feito pela CAPES. Entretanto, diversas métricas capturaram a tendência geral do ranqueamento feito pela CAPES. Por exemplo:

- Em geral, o grau médio do subgrafo induzido apresenta maiores valores para os programas com nível de excelência de acordo com a CAPES. No entanto, há divergências no ranqueamento dos mesmos, por exemplo, o 7A, programa com nota CAPES igual a 7, possui o grau médio do subgrafo induzido menor do que todos os programas com nota CAPES igual a 6. Quanto à classificação dos programas de níveis 3, 4 e 5, existe uma tendência em ter um ranqueamento semelhante ao da CAPES;

- Em geral, o grau médio dos vértices dos programas com nível de excelência de acordo com a CAPES são maiores que os de nível mais baixo, porém, o programa 4B (de nota 4) supera o 6B (de nota 6);

- Os programas com os maiores desvios-padrões apresentados nas tabelas são, em geral, programas melhores classificados de acordo com a CAPES, logo, foi possível verificar que nos programas de qualidade, há uma grande variação entre os pesos dos cortes, enquanto que a maioria dos programas de menor qualidade têm uma pequena variação nestes pesos;

- O tamanho da maior componente conexa do subgrafo induzido normalizado pelo número de vértices possui tendência em classificar vários programas nos níveis semelhantes ao da CAPES, porém os programas 4B e 4F, de nota 4, têm valores semelhantes aos dos programas de excelência;

- O número médio de publicações por vértice, o peso médio dos vértices e o peso médio do corte a partir dos vértices de dentro do conjunto separam melhor os programas de excelência dos demais, no entanto, há muitos programas de nota 4 com médias menores que programas de nota 3.

Algumas métricas não capturaram a tendência geral do ranqueamento feito pela CAPES, como por exemplo, o coeficiente de Gini e o número de componentes conexas do subgrafo induzido. O coeficiente de Gini não é uma boa métrica objetiva para ranqueamento dos programas, pois não há uma correlação entre o ranqueamento da CAPES com a classificação feita pelo coeficiente de Gini, logo, a desigualdade da distribuição do peso do corte dos vértices não é uma boa métrica para o ranqueamento.

Com o intuito de identificar uma maior correlação entre as métricas apresentadas e a classificação subjetiva feita pela CAPES, utilizaram-se os dados das tabelas 4.1 e 4.2 para gerar gráficos com valores de diferentes métricas em cada eixo, onde cada ponto corresponde à um programa de pós-graduação e seu formato corresponde à nota dada pela CAPES em sua avaliação, de acordo com a legenda.

A figura 4.2, apresenta o número médio de publicações e o peso médio dos vértices do programa. Pode-se observar que os programas de excelência (representados por

círculos amarelos e triângulos azuis) se destacam dos demais. Por exemplo, todos os programas com x maior que 18 e y maior que 21 são programas de excelência. Entretanto, esta avaliação não faz uma boa diferenciação dos programas de níveis mais baixos, pois os programas com x entre 10 e 18 e y entre 10 e 21 são de níveis 3 (representados por asteriscos azuis), 4 (representados por quadrados contornados de cor rosa) e 5 (representados por quadrados preenchidos de cor verde). pdf

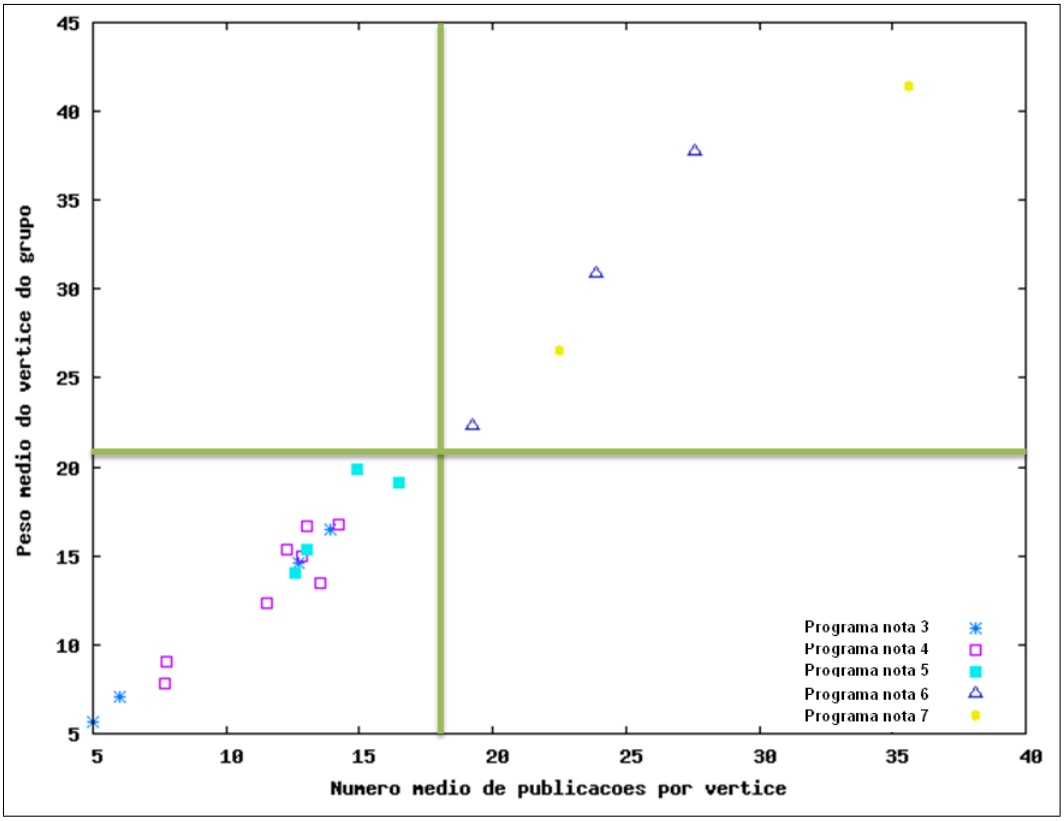


Figura 4.2: Número médio de publicações por vértice x peso médio dos vértices do programa).

A figura 4.3, que utiliza a mesma legenda da figura 4.2, representa o peso médio dos vértices e o peso médio dos vértices do programa no corte considerando somente as colaborações com pesquisadores que atuam fora do Brasil. Nota-se que os programas de excelência se destacam dos demais. Por exemplo, todos os programas com x maior que 17 e y maior que 9,5 são programas de excelência. Em geral os programas com notas 4 e 5 também destacam-se mais que os programas de nota 3. Por exemplo, os programas com x menor que 17 e y entre 5,8 e 9,5 são apenas de níveis 4 e 5. Entretanto, alguns programas de nota 4 estão juntos de programas de nota 3 para x menor que 17 e y menor que 5,8. Portanto, esta é uma boa métrica para identificar programas de qualidade, porém não é capaz de reproduzir fielmente o ranqueamento da CAPES.

Já que as métricas dos pesos médios dos vértices no corte combinadas, como nas figuras 4.2 e 4.3, fazem uma classificação semelhante à avaliação subjetiva da

CAPES, pode-se concluir que a métrica baseada em intensidade do relacionamento entre conjuntos de vértices é uma boa indicação de qualidade. Ou seja, programas de pós-graduação que tem uma média de intensidade de relacionamento alta com pesquisadores de fora do programa e também com pesquisadores do exterior, tendem a ter alta qualidade segundo o ranqueamento da CAPES.

Estas métricas identificam subconjuntos que podem ser utilizados para um processo de avaliação, pois as métricas ranqueiam os mesmos, não necessariamente respeitando uma ordenação de qualidade, mas existe uma correlação muito forte.

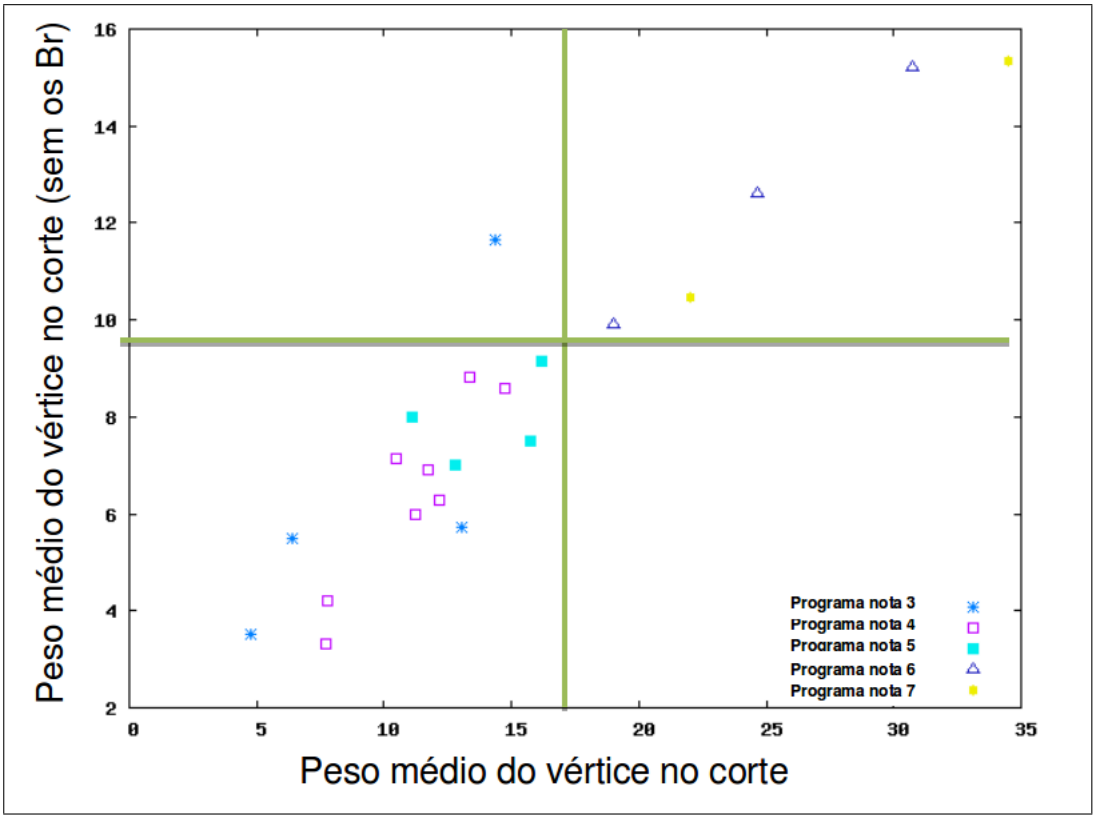


Figura 4.3: Peso médio dos vértices no corte x peso médio dos vértices no corte que atuam fora do Brasil).

4.3 Ranqueamento dos Pesquisadores que atuam no Brasil

Nesta seção, será feito o ranqueamento dos pesquisadores que atuam no Brasil utilizando quatro métricas objetivas, dentre elas a métrica proposta nesta dissertação. O objetivo é verificar se a métrica proposta é capaz de identificar através de um ranqueamento pesquisadores brasileiros influentes da área de Ciência da Computação.

Para constatar a eficácia das métricas, comparou-se os resultados com uma avaliação subjetiva determinada pelo programa Produtividade em Pesquisa (PQ) do

CNPq, que concede bolsas aos pesquisadores para incentivar a pesquisa no Brasil. É importante ressaltar que a avaliação das bolsas de produtividade em pesquisa é, em parte, uma avaliação subjetiva de indivíduos e não leva nenhum aspecto da rede social de colaboração. Cada um dos pesquisadores que participam do programa PQ do CNPq tem uma classificação que leva em consideração os seguintes itens [40]:

- Produção científica e "qualidade" da produção;
- Formação de recursos humanos em nível de Pós-Graduação;,
- Contribuição científica e tecnológica e para inovação;
- Coordenação ou participação principal em projetos de pesquisa;
- Participação em atividades editoriais e de gestão científica e administração de instituições e núcleos de excelência científica e tecnológica.

A classificação do CNPq é dividida em duas categorias 1 e 2. A categoria 1 é dividida em quatro níveis (A, B, C e D), baseada nos critérios acima. Um pesquisador começa no nível 2 e para passar ao nível 1, precisa de oito anos, no mínimo, desde a conclusão de seu doutorado[40]. O nível 1A é reservado a pesquisadores que tenham mostrado excelência continuada na produção científica e na formação de recursos humanos, e que liderem grupos de pesquisa consolidados.

Na comparação dos resultados, utilizou-se as métricas de precisão e abrangência definidas na seção 2.2 para verificar o grau de acerto de cada uma das quatro métricas de ranqueamento com relação aos pesquisadores classificados de acordo com o CNPq. As quatro métricas utilizadas e aplicadas aos pesquisadores que atuam no Brasil foram: número de publicações, número de colaboradores (grau) dos pesquisadores, peso dos vértices (métrica de Newman) e a métrica proposta por este artigo, o peso do vértice no corte definido pelo conjunto de pesquisadores que atuam no Brasil.

Foram feitos quatro ranqueamentos, um para cada métrica. Por exemplo, na métrica grau, a primeira entrada é o vértice de maior grau, logo este corresponde ao primeiro lugar no ranqueamento. A segunda entrada é o vértice de segundo maior grau, então este corresponde ao segundo lugar no ranqueamento, e assim por diante. E assim, observou-se a distribuição dos bolsistas de produtividade em pesquisa por categoria e nível em cada ranqueamento. Por exemplo, na lista dos 20 pesquisadores com maior grau, existem x pesquisadores de nível 1A, ou seja, x pesquisadores foram identificados. Desta forma pode-se calcular a precisão e abrangência e medir a eficácia da métrica em identificar os pesquisadores de nível 1A.

De posse dos ranqueamentos de cada métrica, considerou-se listas contendo os n pesquisadores mais bem colocados em cada ranqueamento, onde n assumiu os valores de 20, 25, 30, 35 e 50.

A relação de bolsistas de produtividade em pesquisa foi obtida em [33]. Eles foram identificados manualmente na DBLP, de forma semelhante à identificação dos pesquisadores que atuam no Brasil descrita na seção 3.1.

A tabela 4.3 apresenta a abrangência e precisão dos ranqueamentos feitos a partir das quatro métricas aplicadas aos pesquisadores que atuam no Brasil, na tentativa de identificar os pesquisadores com bolsa de produtividade em pesquisa do CNPq de nível 1A, variando-se, de 20 a 50, o tamanho da lista dos primeiros pesquisadores ranqueados. Os valores em negrito correspondem aos maiores valores de precisão e abrangência para cada tamanho de lista. Utilizando as listas de tamanho 20 a 35 elementos, os pesquisadores de nível 1A são melhores recuperados pela métrica peso do vértice no corte, pois é a métrica com maiores valores de precisão e abrangência. Já em uma lista de 50 pesquisadores, as métricas número de publicações, peso do vértice no corte e peso do vértice têm o mesmo desempenho. Podemos concluir que o peso do vértice no corte se mostra a melhor métrica para identificar pesquisadores de nível de excelência.

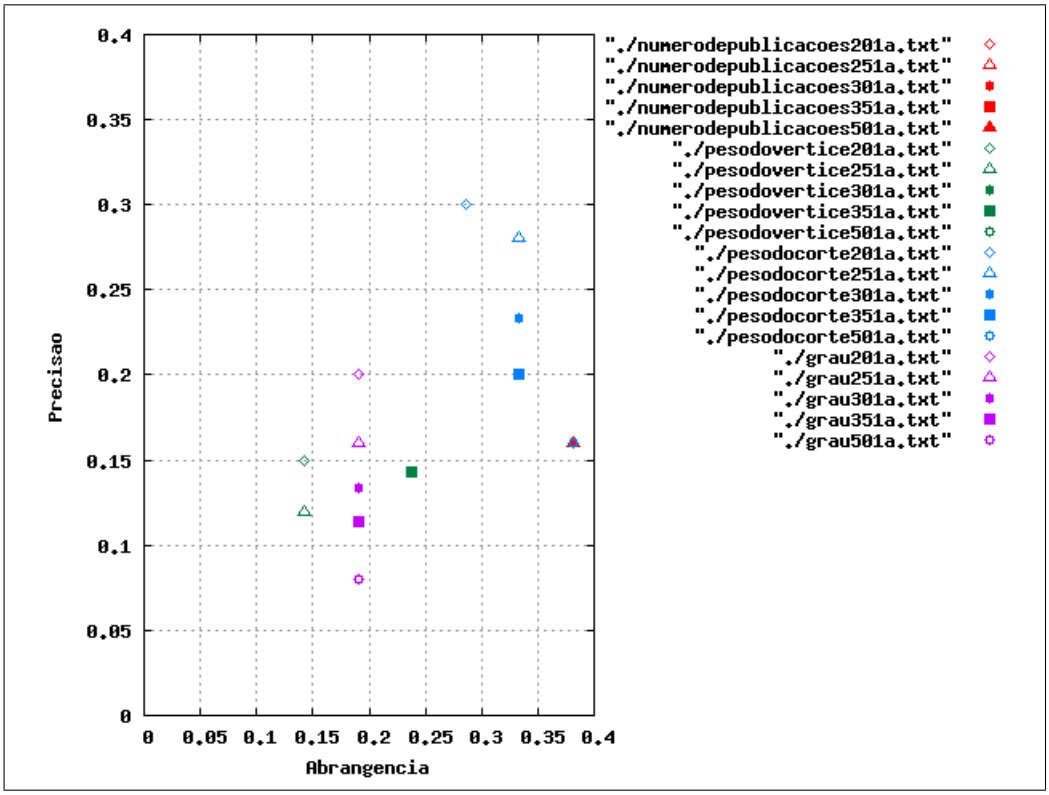


Figura 4.4: Precisão e abrangência dos ranqueamentos utilizando métricas diferentes para retornar pesquisadores com bolsa de produtividade de pesquisa 1A e variando o tamanho da lista.

Uma observação importante acerca da figura 4.4 é que o peso do vértice e o número de publicações possuem o mesmo valor de precisão e abrangência para os tamanhos de conjunto de 20 a 35. Isso se deve à semelhança das duas métricas, pois o peso do vértice corresponde ao número de publicações com ao menos uma

Tabela 4.3: Dados utilizados para plotar a figura 4.4, contendo a precisão e abrangência ao retornar pesquisadores de nível 1A dos ranqueamentos feitos através de diferentes métricas objetivas.

Tamanho da Lista	Número de Publicações		Peso do Vértice		Peso do Vértice no Corte		Grau	
	Precisão	Abrangência	Precisão	Abrangência	Precisão	Abrangência	Precisão	Abrangência
20	0,15	0,14286	0,15	0,14286	**0,3**	**0,2857**	0,2	0,1905
25	0,16	0,19048	0,12	0,14286	**0,28**	**0,3333**	0,16	0,1905
30	0,13	0,19048	0,13	0,19048	**0,23**	**0,3333**	0,13	0,1905
35	0,14	0,2381	0,14	0,2381	**0,2**	**0,3333**	0,11	0,1905
50	**0,16**	**0,38095**	**0,16**	**0,38095**	**0,16**	**0,381**	0,08	0,1905

colaboração. Pode-se induzir a partir desta observação, que os pesquisadores de nível 1A, em geral, publicam em colaboração.

A figura 4.5 apresenta um gráfico semelhante ao da figura 4.4, com apenas uma diferença, aumentou-se o número de objetos relevantes, ou seja, considerou-se pesquisadores 1A e 1B como sendo relevantes. A tabela 4.4 contém os dados utilizados para gerar o gráfico. Os valores em negrito correspondem aos maiores valores de precisão e abrangência para cada tamanho de lista. Devido ao aumento do número de objetos relevantes, a precisão aumentou em todas as métricas. Logo, pode-se concluir que as métricas peso do vértice, número de publicações e grau, têm um bom desempenho ao aumentar o número de objetos relevantes, ou seja, capturam bem os pesquisadores de níveis 1A e 1B. As métricas número de publicações e peso do vértice tiveram melhor desempenho ao utilizar listas de tamanho 35 e 50.

Tabela 4.4: Dados utilizadados para plotar a figura 4.5, contendo a precisão e abrangência ao retornar pesquisadores de níveis 1A e 1B dos ranqueamentos feitos através de diferentes métricas objetivas.

Tamanho da Lista	Número de Publicações		Peso do Vértice		Peso do Vértice no Corte		Grau	
	Precisão	Abrangência	Precisão	Abrangência	Precisão	Abrangência	Precisão	Abrangência
20	0.25	0.12	0.25	0.12	**0.45**	**0.21**	0.35	0.17
25	0.36	0.21	0.24	0.14	**0.4**	**0.24**	0.36	0.21
30	0.3	0.21	0.3	0.21	**0.33**	**0.24**	0.3	0.21
35	**0.31**	**0.26**	**0.31**	**0.26**	0.29	0.24	0.26	0.21
50	**0.3**	**0.36**	**0.3**	**0.36**	0.26	0.31	0.22	0.26

A figura 4.6 mostra o comportamento da métrica peso do corte ao variar o tamanho da lista e o número de objetos relevantes, onde a variação da cor corresponde ao aumento do número de objetos relevantes e a variação da forma corresponde ao aumento do número de objetos recuperados. Através do gráfico comprova-se que ao aumentar o número de objetos relevantes, a abrangência diminui e a precisão aumenta. Já ao aumentar o número de objetos recuperados a abrangência aumenta e a precisão diminui.

Uma comparação direta entre as métricas pode ser vista na figura 4.7 que apre-

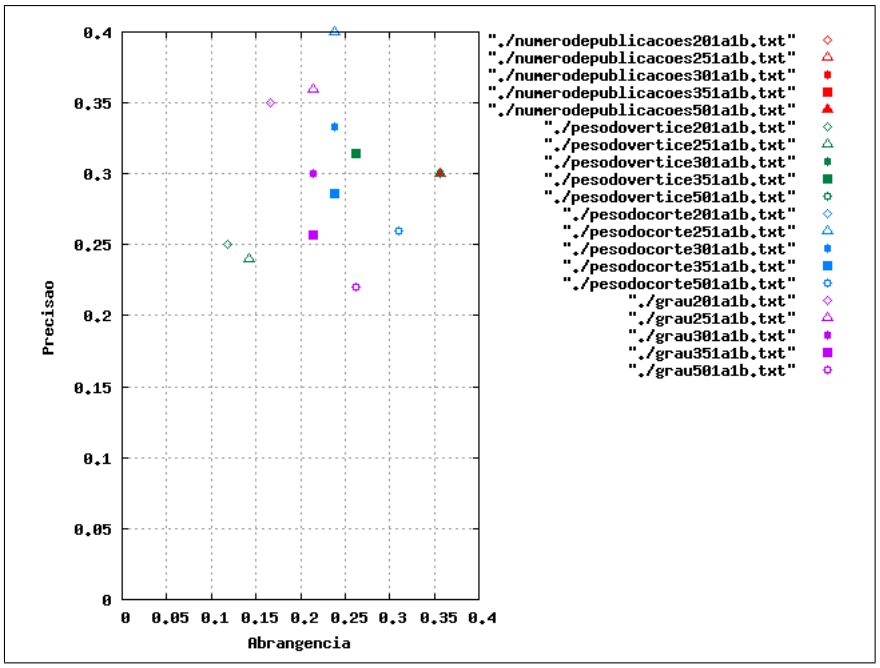


Figura 4.5: Precisão e abrangência dos ranqueamentos utilizando métricas diferentes para retornar pesquisadores com bolsa de produtividade de pesquisa 1A e 1B e variando o tamanho da lista.

senta um gráfico em barras, com o valor da medida-F para as diferentes métricas. O peso do vértice no corte contém uma medida-F superior às demais métricas ao retornar pesquisadores de nível 1A em uma lista de 20 indivíduos. Ao aumentar o tamanho da lista para 25 e mantendo o conjunto de pesquisadores 1A, o desempenho da métrica também aumenta em relação às outras. Para uma lista com 20 elementos, quando aumentamos o conjunto de objetos relevantes, ou seja, considerando pesquisadores 1A e 1B como relevantes, a medida-F do peso do vértice no corte tem uma pequena queda em seu valor e as demais métricas aumentam seu desempenho. Entretanto, o peso do vértice no corte continua tendo um valor superior a todos as outras métricas. E, finalmente, ao aumentar o tamanho da lista para 25 e mantendo o conjunto 1A e 1B como relevantes, o maior desempenho é do peso do vértice no corte, porém as métricas grau e número de publicações se mostram muito eficientes em recuperar pesquisadores de nível 1B, pois há um grande aumento na medida-F destas métricas ao acrescentar os pesquisadores 1B na lista de objetos relevantes.

A tabela 4.5 mostra o ranqueamento dos pesquisadores de nível 1A nas quatro métricas avaliadas. Os nomes dos pesquisadores estão em anonimato e os valores em negrito indicam a métrica que melhor recupera o pesquisador correspondente, ou seja, a métrica em que o indivíduo tem a maior posição. A métrica peso do corte no vértice é a que possui maior número de valores em negrito, logo ela é a que melhor classifica pesquisadores de nível 1A ao realizar um ranqueamento de todos os pesquisadores que atuam no Brasil. Em seguida vem a métrica peso do vértice,

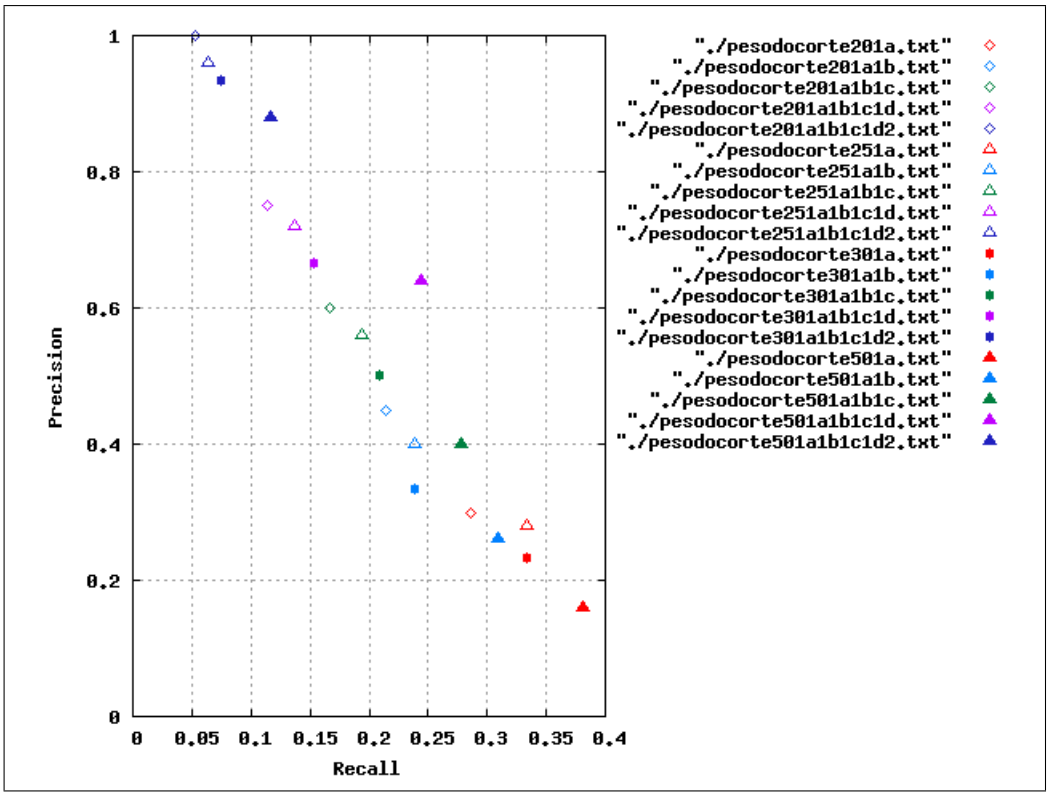


Figura 4.6: Precisão e abrangência aplicadas à métrica peso do corte, variando o tamanho do conjunto e o número de objetos relevantes.

que possui sete valores em negrito. Pode-se observar que não há uma correlação entre as posições dos pesquisadores na métrica peso do vértice no corte e as demais. Por exemplo, enquanto o pesquisador EE está na oitava posição através da métrica peso do vértice no corte, na métrica grau está na octogésima quarta posição e na quinquagésima quarta no número de publicações. O pesquisador BB, segundo lugar na métrica peso do corte, está na primeira posição nas demais métricas.

A tabela 4.6 apresenta a classificação dos pesquisadores da rede brasileira através da métrica peso do vértice no corte, na qual os nomes dos pesquisadores estão mantidos em anonimato. Dos 25 indivíduos recuperados, 24 são bolsistas de produtividade em pesquisa, considerando todos os níveis. Destes, sete possuem nível 1A (pesquisadores representados por letras iguais), três são 1B, seis são 1C, três são 1D e cinco possuem nível 2. As posições dos indivíduos em outras métricas também são mostradas. O primeiro lugar ficou com AB, bolsista de produtividade em pesquisa do CNPq nível 1C. Nas métricas peso do vértice, número de publicações e grau, suas posições foram 17, 16 e 23, respectivamente.

Resumindo, a métrica proposta foi a que melhor classificou os bolsistas de produtividade em pesquisa. Logo, é uma boa métrica para identificar bons pesquisadores em redes de colaboração quando tratamos de pesquisadores 1A e 1A/1B.

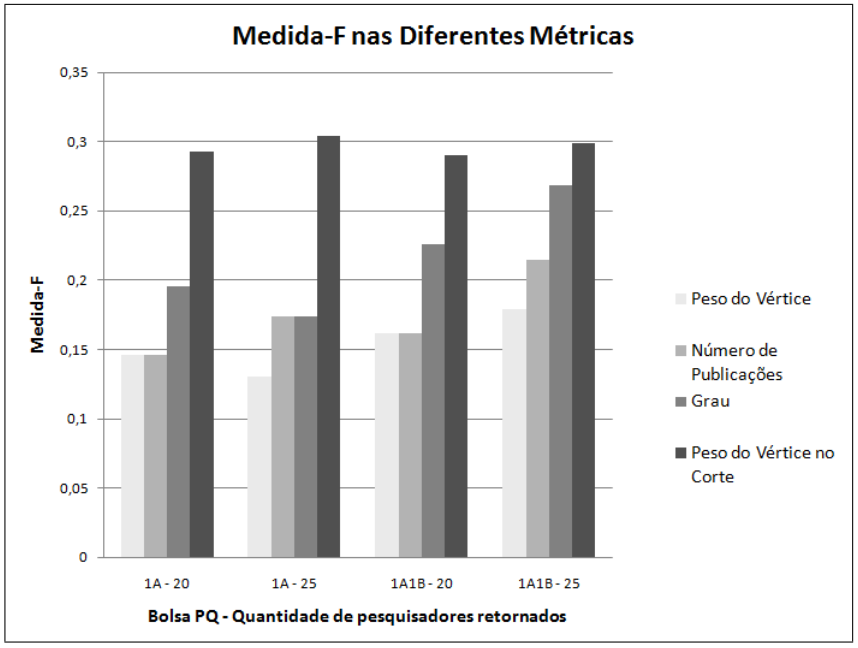


Figura 4.7: Medida-F aplicada nas quatro métricas de ranqueamento ao retornar pesquisadores de nível 1A e 1B.

Tabela 4.5: Classificação dos pesquisadores de nível 1A nas quatro métricas avaliadas. Cada célula da tabela possui a posição do pesquisador na classificação utilizando cada métrica.

NOME	PESO DO VERTICE NO CORTE	PESO DO VERTICE	GRAU	NUM. DE PUBLIC.
AA	23	**3**	15	**3**
BB	2	**1**	**1**	**1**
CC	41	**32**	67	35
DD	**219**	368	651	349
EE	**8**	51	84	54
FF	210	141	155	**119**
GG	**17**	29	63	19
HH	**201**	235	787	241
II	212	174	**173**	180
JJ	164	**103**	118	111
LL	208	163	**97**	169
MM	**244**	282	711	290
NN	99	**41**	71	44
OO	283	90	352	**72**
PP	411	**333**	1530	341
QQ	**61**	102	242	106
RR	139	**119**	484	125
SS	**18**	36	104	37
TT	20	20	**18**	21
UU	**7**	47	91	50
VV	169	**144**	309	149

Tabela 4.6: Classificação dos pesquisadores da rede brasileira através da métrica proposta neste trabalho.

NOME	BOLSA PQ	PESO DO VERTICE NO CORTE	PESO DO VERTICE	NUMERO DE PUBLICAÇOES	GRAU
AB	1C	1º	17º	16º	23º
BB	1A	2º	1º	1º	1º
AC	1D	3º	43º	41º	40º
AD	2	4º	57º	61º	301º
AE	2	5º	13º	15º	10º
AF	1C	6º	16º	18º	14º
UU	1A	7º	47º	50º	91º
EE	1A	8º	51º	54º	84º
AH	1B	9º	25º	25º	8º
AI	1C	10º	58º	51º	77º
AJ	1B	11º	28º	23º	36º
AL	2	12º	46º	49º	96º
AM	1D	13º	4º	4º	2º
AN	2	14º	21º	24º	39º
AO	1C	15º	37º	39º	41º
AP	2	16º	65º	67º	52º
GG	1A	17º	29º	19º	63º
SS	1A	18º	36º	37º	104º
AQ	1C	19º	12º	13º	27º
TT	1A	20º	20º	21º	18º
AR	1C	21º	7º	7º	16º
AS	1B	22º	52º	32º	35º
AA	1A	23º	3º	3º	15º
AT	--	24º	42º	47º	139º
AU	1D	25º	27º	31º	62º

Capítulo 5

Conclusão e Trabalhos Futuros

5.1 Conclusões

Dentre os vários tipos de redes, existem as redes sociais que possuem grande importância por representar a interação humana. As características topológicas dessas redes refletem o comportamento social de seus participantes. Dentre as redes sociais, existem as redes de colaboração científica, que vêm sendo estudadas por muitos pesquisadores, com o intuito não apenas de entender as características dessas redes, mas também poder criar novos serviços. Intuitivamente, relacionamentos em redes de colaboração possuem diferentes intensidades. Para medir essa intensidade, pode-se utilizar uma métrica baseada na quantidade de colaborações e número de publicações. Através dessas métricas, pode-se definir importância de indivíduos ou conjuntos de indivíduos, através de um ranqueamento dos mesmos segundo esta métrica.

A contribuição principal desse trabalho é o desenvolvimento de uma métrica para ranqueamento em redes de colaboração baseada na intensidade de relacionamento entre indivíduos e conjuntos de indivíduos. Esta métrica é baseada na proposta de Newman [6], mas utiliza-se do conceito de corte em grafos. Sua fundamentação vem da ideia de que é possível medir relevância de subconjunto de indivíduos para a rede, ou a importância de um indivíduo para um grupo de indivíduos, através da relação existente entre as intensidades dos relacionamentos.

Este trabalho também apresenta um estudo de diversas propriedades topológicas da rede de colaboração mundial e da rede de colaboração brasileira de autores de artigos científicos da área de Ciência da Computação. Além disto, apresenta a comparação dos ranqueamentos feitos a partir de métricas objetivas com os ranqueamentos subjetivos feitos por órgãos de grande credibilidade na área acadêmica do Brasil. Dessa forma, os principais pesquisadores e grupos que atuam no Brasil puderam ser identificados por métricas topológicas.

Ao estudar as propriedades topológicas da rede de colaboração mundial e da rede de colaboração brasileira, obteve-se as seguintes conclusões:

- Existe uma correlação entre o número de publicações e o número de colaboradores, pois, em geral, há uma tendência a quanto maior o número de publicações, maior o número de co-autores;
- Os pesquisadores que atuam no Brasil têm, em média, mais publicações em colaboração do que o restante do mundo, porém estes tendem a ter um menor número de colaboradores;
- A rede brasileira é mais nova nova do que a rede mundial, pois tanto as publicações quanto os vértices são mais recentes.

Ao identificar os programas de pós-graduação da área de Ciência da Computação do Brasil através de métricas topológicas, concluiu-se que nenhuma métrica objetiva aplicada à rede de colaboração foi capaz de reproduzir o ranqueamento subjetivo feito pela CAPES. Entretanto, diversas métricas capturaram a tendência geral do ranqueamento feito pela CAPES. Por exemplo, programas de pós-graduação que têm uma média de intensidade de relacionamento alta com pesquisadores de fora do programa e também com pesquisadores do exterior, tendem a ter alta qualidade.

Por fim, através de experimentações empíricas, concluiu-se que:

- A métrica proposta foi a que apresentou melhor desempenho entre as métricas objetivas ao recuperar pesquisadores de níveis 1A e 1A/1B em listas de tamanho 20 e 25;
- A métrica proposta foi a que melhor classificou os pesquisadores de nível 1A ao realizar o ranqueamento de todos os pesquisadores que atuam no Brasil.

5.2 Trabalhos Futuros

Durante o desenvolvimento desse trabalho algumas ideias interessantes surgiram, sem, no entanto, haver tempo hábil para investigá-las. Abaixo, lista-se algumas destas ideias:

- Adaptar a métrica proposta para torná-la uma métrica temporal. Ou seja, considerar a idade da interação (idade da publicação) na definição da intensidade de relacionamento. Essa ideia surgiu a partir da intuição de que colaborações antigas tem menor importância na intensidade do relacionamento entre dois indivíduos;

- Utilizar a métrica propostas para avaliar um modelo de propagação de informação. A ideia se baseia no fato de que informação se propaga mais rapidamente por relacionamentos de maior intensidade;
- Definir uma métrica recursiva de importância que leva em consideração não só os vizinhos do vértice, mas os vizinhos dos vizinhos, os vizinhos dos vizinhos dos vizinhos, e assim por diante;
- Incluir na métrica o fator de impacto da publicação de acordo com a importância da conferência ou periódico.

Referências Bibliográficas

[1] KLEINBERG, J., TARDOS, É. *Algorithm desing.* Pearson Education, 2006. ISBN: 0-321-29535-8.

[2] NEWMAN, M. E. J. "The structure of scientific collaboration networks". In: *Proc. Natl. Acad. Sci. USA*, v. 98, pp. 404–409, jan. 2001.

[3] NEWMAN, M. E. J. "The Structure and Function of Complex Networks", *SIAM Review*, v. 45, pp. 167–256, 2003.

[4] ONODY, R. N., DE CASTRO, P. A. *Complex network study of Brazilian soccer players*, Oct 2004.

[5] BARABASI, A.-L. *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life.* Plume Books, April 2003. ISBN: 0452284392.

[6] NEWMAN, M. E. J. "Who Is the Best Connected Scientist? A Study of Scientific Coauthorship Networks". v. 650/2004, *Lecture Notes in Physics*, pp. 337–370, Springer Berlin / Heidelberg, ago. 2004. ISBN: 978-3-540-22354-2. doi: 10.1007/b98716.

[7] WAGNER, C., LEYDESDORFF, L. "Network structure, self-organization, and the growth of international collaboration in science", *Research Policy*, v. 34, n. 10, pp. 1608–1618, December 2005. ISSN: 00487333. Disponível em: <http://dx.doi.org/10.1016/j.respol.2005.08.002>.

[8] DE ANDRADE MENEZES, V. S., DA SILVA, R. T., DE SOUZA, M. F., etal. "Mining and Analyzing Organizational Social Networks Using Minimum Spanning Tree". In: *OTM '08: Proceedings of the OTM Confederated International Workshops and Posters on On the Move to Meaningful Internet Systems*, pp. 18–19, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN: 978-3-540-88874-1. doi: http://dx.doi.org/10.1007/978-3-540-88875-8_9.

[9] BARCELLOS, C., BASTOS, F. I. "Redes sociais e difusão da AIDS no Brasil", *Organización Panamericana de la Salud*, 1996. ISSN: 1020-4989.

[10] HAYASHI, M. C. P. I., HAYASHI, C. R. M., DE LIMA, M. Y. "Análise de redes de co-autoria na produção científica em educação especial", *Liinc em Revista*, v. 4, n. 1, pp. 84–103, mar. 2008.

[11] OLIVEIRA, E. F. T., DA SILVA SANTAREM, L. G., SEGUNDO, J. E. S. "Análise das redes de colaboração científica através do estudo das co-autorias, nos cursos de pós-graduação do brasil no tema tratamento temático da informação", *Nuevas perspectivas para la difusión y organización del conocimiento: actas del congreso / coord. por Nuria Lloret Romero*, v. 2, pp. 986–1000, set. 2009.

[12] GROSSMAN, J., ION, P. *The Erdös Number Project*. Disponível em: <`http://www.oakland.edu/enp/`>.

[13] REYNOLDS, P. *The Oracle of Bacon*. Disponível em: <`http://oracleofbacon.org`>.

[14] NEWMAN, M. E. J. "Coauthorship networks and patterns of scientific collaboration". In: *Proc. Natl. Acad. Sci. USA*, v. 101, pp. 5200–5205, abr. 2004. doi: 10.1073/pnas.0307545100.

[15] MENEZES, G. V., ZIVIANI, N., LAENDER, A. H. F., etal. "A geographical analysis of knowledge production in computer science". In: *Proceedings of the 18th international conference on World wide web*, pp. 1041–1050, 2009.

[16] HUANG, J., ZHUANG, Z., LI, J., etal. "Collaboration over time: characterizing and modeling network evolution". In: *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pp. 107–116, New York, NY, USA, 2008. ACM. ISBN: 978-1-59593-927-9. doi: http://doi.acm.org/10.1145/1341531.1341548.

[17] FREITAS, C. M. D. S., NEDEL, L. P., GALANTE, R., etal. "Extração de Conhecimento e Análise Visual de Redes Sociais". In: *SEMISH (Seminário Integrado de Software e Hardware)*, pp. 106–120, 2008.

[18] GLOOR, P. A., LAUBACHER, R., DYNES, S. B. C., etal. "Visualization of Communication Patterns in Collaborative Innovation Networks - Analysis of Some W3C Working Groups". In: *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pp. 56–60, New York, NY, USA, 2003. ACM. ISBN: 1-58113-723-0. doi: http://doi.acm.org/10.1145/956863.956875.

[19] DE NOOY, W., MRVAR, A., BATAGELJ, V. *Exploratory Social Network Analysis with Pajek (Structural Analysis in the Social Sciences)*. Cambridge University Press, 2005. ISBN: 0521602629.

[20] MEDEIROS, M. "Uma introdução às representações gráficas da desigualdade de renda", *IPEA*, 2006. ISSN: 1415-4765.

[21] DE JESUS DE SOUZA, N. *Uma introdução à história do pensamento economico*. Disponível em: <http://www.nalijsouza.web.br.com/introd_hpe.pdf>.

[22] MARQUES, P. "Contribuição ao estudo da organização agroindustrial: o caso da indústria de frango de corte no Estado de São Paulo". In: *Scientia Agricola*, v. 51, abr. 1994. doi: 10.1590/S0103-90161994000100002.

[23] *Human Development Report 2009 - Economy and inequality - Gini Index*. Disponível em: <http://hdrstats.undp.org/en/indicators/161.htmlMEco>.

[24] *File:Gini Coefficient World Human Development Report 2007-2008.png*. Disponível em: <http://commons.wikimedia.org/wiki/File:Gini_Coefficient_World_Human_Development_Report_2007-2008.png>.

[25] BELKIN, N. J., CROFT, W. B. "Information filtering and information retrieval: two sides of the same coin?" *Commun. ACM*, v. 35, n. 12, pp. 29–38, 1992. ISSN: 0001-0782. doi: http://doi.acm.org/10.1145/138859.138861.

[26] ZHU, M. *Recall, Precision and Average Precision*, ago. 2004. Disponível em: <http://www.stats.uwaterloo.ca/stats_navigation/techreports/04WorkingPapers/2004-09.pdf>.

[27] KANDEFER, M., SHAPIRO, S. "An F-Measure for Context-Based Information Retrieval". In: *Commonsense 2009: the Ninth International Symposium on Logical Formalizations of Commonsense Reasoning*. The Fields Institute, jun. 2009.

[28] *A Plataforma Lattes*. Disponível em: <http://lattes.cnpq.br/>.

[29] *The DBLP Computer Science Bibliography*. Disponível em: <http://www.informatik.uni-trier.de/~ley/db/>.

[30] LEY, M. "DBLP — Some Lessons Learned", *VLDB '09*, 2009.

[31] *DBLP FAQ: How to parse dblp.xml?* Disponível em: <http://www.informatik.uni-trier.de/~ley/db/about/simpleparser/index.html>.

[32] *Docentes da pós-graduação e bolsistas de produtividade em pesquisa.* Disponível em: <http://ged.capes.gov.br/AgProd/silverstream/pages/pgRelBolsistasProdPesqResultado.html>.

[33] CNPQ. *Bolsas de Produtividade em Pesquisa - Bolsas em curso.* Disponível em: <http://plsql1.cnpq.br/divulg/RESULTADO_PQ_102003.curso>.

[34] BEZERRA, R. L. *Análise da Conectividade em Redes Móveis Utilizando Dados Obtidos da Mobilidade Humana.* Tese de Mestrado, Universidade Federal do Rio de Janeiro/COPPE, mar. 2009.

[35] ALBERT, R., BARABÁSI, A.-L. "Statistical mechanics of complex networks", *CoRR*, v. cond-mat/0106096, 2001.

[36] TRAVERS, J., MILGRAM, S. "An Experimental Study of the Small World Problem", *Sociometry*, v. 32, pp. 425–443, 1969.

[37] SZWARCFITER, J. L. *Grafos e algoritmos computacionais.* Campus, 1986.

[38] *Docentes da pós-graduação e bolsistas de produtividade em pesquisa.* Disponível em: <http://ged.capes.gov.br/AgProd/silverstream/pages/pgRelBolsistasProdPesq.html>.

[39] *Avaliação da pós-graduação.* Disponível em: <http://www.capes.gov.br/avaliacao/avaliacao-da-pos-graduacao>.

[40] *CNPq - Instrumentos Normativos - Produtividade em Pesquisa (PQ).* Disponível em: <http://universia.com.br/materia/materia.jsp?id=7585>.