Ranqueamento em Redes de Colaboração Utilizando uma Métrica Baseada em Intensidade do Relacionamento

Vinícius P. de M. Freire, Daniel R. Figueiredo
COPPE/Programa de Engenharia de Sistemas e Computação (PESC)
Universidade Federal do Rio de Janeiro (UFRJ)
Caixa Postal 68511 - Rio de Janeiro, RJ - 21941-972
{vini,daniel}@land.ufrj.br

***Abstract*—Collaboration networks are social networks in which relationships represent some kind of professional collaboration. The study of collaboration networks can help identify members that are important or influential within a given community. 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. The second part presents a ranking metric for vertices based on the intensities of their relationships. Using the proposed metric and other more classical metrics, we rank the researchers in Brazil in Computer Science. The proposed metric was compared with subjective evaluation of researchers made by CNPq. The results show the effectiveness of the proposed metric in identifying influential researchers when compared to another metrics in the literature.**

***Keywords*-coauthorship networks; intensity of relationship; ranking metric;**

***Resumo*—Redes de colaboração são redes sociais nas quais os relacionamentos entre indivíduos representam algum tipo de colaboração profissional. O estudo de redes de colaboração pode ajudar a identificar indivíduos que sejam influentes ou importantes dentro de uma determinada comunidade. 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. A segunda parte apresenta uma métrica para ranqueamento de vértices baseada na intensidade de relacionamento. Utilizando a métrica proposta e outras métricas clássicas, fazemos um ranqueamento dos pesquisadores que atuam no Brasil na área de Ciência da Computação. A avaliação dos resultados foi feita através da comparação com a avaliação subjetiva de pesquisadores feita pelo CNPq. Os resultados evidenciam a eficácia da métrica em identificar indivíduos influentes quando comparada a outras métricas presentes na literatura.**

***Palavras-chave*-redes de colaboração; intensidade de relacionamento; métrica para ranqueamento;**

I. Introdução

Redes sociais são importantes abstrações que vem sendo estudadas há décadas por diversas áreas do conhecimento. Uma rede social é formada por um conjunto de indivíduos e por relacionamentos entre pares de indivíduos [1]. Uma rede social pode capturar diferentes tipos relacionamentos tais como amizade, parentesco ou colaboração profissional.

Intuitivamente, relacionamentos podem ter diferentes intensidades. Em redes sociais de amizade, por exemplo, existem pessoas mais amigas que outras. Em redes de colaboração profissional, existem pessoas que trabalham mais em conjunto do que outras. Portanto, para medir a intensidade dos relacionamentos, ou peso das arestas, devemos utilizar alguma métrica adequada, pois existem diferentes formas de medir intensidade de relacionamento.

Uma rede de colaboração científica é uma rede social onde os vértices são autores de artigos científicos, e existe uma aresta entre dois autores se eles publicaram juntos, ou seja, colaboraram na produção de um ou mais artigos científicos. 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 pesquisadores tiverem em conjunto, maior será a intensidade do relacionamento, ou peso da aresta.

Ao analisar uma rede de colaboração podem-se descobrir muitas de suas propriedades topológicas. Estas características podem ser utilizadas para realizar 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 [1] [2] [3].

Logo, a motivação deste artigo é utilizar redes de colaboração e intensidade de relacionamento para definir uma métrica objetiva que captura importância de indivíduos. Assim, através desta métrica, podemos fazer um ranqueamento dos indivíduos de um subconjunto da

rede de colaboração.

A contribuição deste artigo se divide em duas partes. A primeira delas é um estudo de algumas 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 e dos vértices. A segunda parte, e mais importante, é o desenvolvimento de uma métrica baseada em intensidade de relacionamento para medir a relevância de vértices em redes de colaboração. Esta métrica permite fazer o ranqueamento de indivíduos dentro de um determinado subconjunto. Para avaliar sua eficácia, realizamos um ranqueamento dos pesquisadores que atuam no Brasil na área da Ciência da Computação e o comparamos com ranqueamentos gerados a partir de outras métricas objetivas. Comparamos estes resultados com o ranqueamento subjetivo feito por um órgão de grande credibilidade na área acadêmica do Brasil (CNPq). Os resultados indicam que a métrica proposta tem um desempenho melhor do que outras métricas objetivas, no sentido de melhor identificar pesquisadores mais influentes de acordo com a avaliação subjetiva.

Este artigo está organizado da seguinte forma. Na seção II nós apresentamos os trabalhos relacionados e algumas discussões. A seção III apresenta a caracterização da rede de colaboração estudada e a análise das propriedades topológicas da rede mundial e da rede formada somente por pesquisadores que atuam no Brasil. A seção IV introduz uma métrica para ranqueamento baseada em intensidade de relacionamento. Na seção IV apresentamos um ranqueamento dos pesquisadores que atuam no Brasil através da métrica proposta e avaliamos sua eficácia através da comparação de seus resultados com a avaliação subjetiva de pesquisadores feita pelo CNPq. Finalmente, na seção V são feitas as considerações finais sobre este estudo sugerindo possíveis temas para trabalhos futuros.

II. Trabalhos Relacionados

Diversas áreas do conhecimento vêm estudando redes sociais com o objetivo de entender diferentes fenômenos [4] [5] [6] [7] [8]. 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, pesquisadores das diversas áreas vêm desenvolvendo metodologias de análise que tem como base as relações entre os indivíduos, considerando uma estrutura em forma de redes [6].

Neste artigo vamos abordar as redes sociais de colaboração científica. 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 pesquisador [7]. O número de Erdös representa a distância na rede de colaboração entre uma pessoa e Paul Erdös, medida pela autoria de trabalhos científicos.

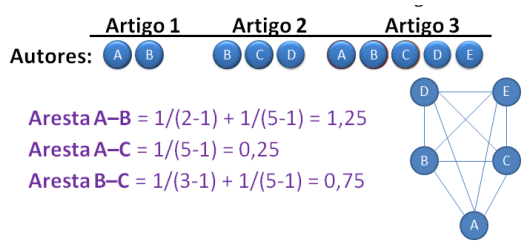

Figura 1. Cálculo dos pesos das arestas utilizando o método de Newman.

Em [9], Newman procura responder a uma variedade de questões sobre os padrões de colaboração científica nas áreas de Biologia, Física e Matemática através do estudo das propriedades topológicas das respectivas redes de colaboração. Menezes et. al analisaram a produção científica em Ciência da Computação do Brasil, América do Norte e Europa, por meio de redes de colaboração, bem como a evolução temporal dessas redes [10].

Em [4], os autores utilizaram um método de detecção de grupos para identificar comunidades de pesquisa na rede social científica brasileira. 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.

Em [8], 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 eficazmente futuras colaborações entre indivíduos baseada na estrutura da vizinhança local.

Em [3], Newman introduziu uma nova métrica para medir a intensidade do relacionamento em redes de colaboração científica, daqui para frente chamada de **Métrica de Newman**. Ela funciona da seguinte maneira: cada artigo escrito 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.

Um exemplo desta métrica está na ilustrado na figura 1, onde existem três artigos escritos pelos autores A, B, C, D e E, cada um com diferentes colaborações. 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.

Newman descreve seu método como sendo o quanto os autores colaboram em cada artigo que eles trabalham juntos. A idéia é que cada artigo tem intensidade constante, que é dividida igualmente entre os co-autores.

Newman definiu o peso do vértice como sendo a soma dos pesos das arestas que incidem no vértice, o que corresponde ao número de artigos que o autor possui em colaboração com algum outro pesquisador.

A métrica de Newman e o peso dos vértices definidos por ele foram utilizados neste artigo para calcular, respectivamente, o peso das arestas e dos vértices.

A diferença entre a métrica de Newman e a métrica proposta neste artigo é que a métrica proposta considera os relacionamentos entre dois conjuntos diferentes de vértices no cálculo da intensidade de um vértice, enquanto a métrica de Newman considera todos os relacionamentos incidentes sobre um vértice.

III. Caracterização da Rede de Colaboração

A rede de colaboração científica utilizada neste trabalho foi construída utilizando a base *DBLP* (*Digital Bibliography & Library Project*) obtida em junho de 2009 [11]. A *DBLP* é uma base de dados [1] 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 [12]. A *DBLP* não é uma base completa da Ciência da Computação, muito menos cataloga artigos publicados em todos os simpósios brasileiros, no entanto pode ser utilizada para ranquear os pesquisadores que atuam no Brasil. Uma vantagem é que nesta base muitos autores com mais de um nome são representados como a mesma pessoa.

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 na seção II.

A. Análise da Rede Mundial

De início fizemos a análise das propriedades estruturais da rede de colaboração mundial que possui 722.392 vértices e 2.272.540 arestas.

Grau

O grau d_i é o número de vizinhos do vértice i, ou seja, o número de colaboradores do autor i.

[1] disponível publicamente em [11].

A distribuição empírica do grau é apresentada na figura 2, onde o eixo x representa o grau dos vértices, ou seja, o número de colaboradores dos autores e o eixo y é a fração dos vértices com grau maior ou igual a x.

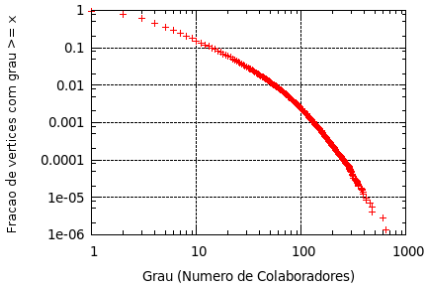

Figura 2. Distribuição do grau dos vértices da rede de colaboração mundial.

Observando a figura 2 percebemos que a distribuição do grau possui uma cauda pesada, ocorrendo valores muito distantes do grau médio, que é 6,3, variando de 0 a 643. Cauda pesada é uma característica comum em redes sociais e redes de colaboração, e foi observada em trabalhos anteriores [9] [13] [2] [8] [10]. Há um pequeno número de autores com muitas colaborações e uma grande quantidade de autores com poucos colaboradores. Por exemplo, no gráfico observamos 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, notamos 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[2].

Peso da Aresta

Calculamos os pesos das arestas da rede de colaboração através da métrica de Newman. Observando a figura 3 percebemos que a distribuição do peso das arestas possui uma cauda pesada, ocorrendo valores muito distantes do peso médio, que é 0,63, variando de 0,0088 a 267,77. Analisando a base DBLP, constatamos 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.

No gráfico, a primeira área sombreada representa pessoas que colaboraram poucas vezes com muitos co-autores. Já segunda área sombreada representa pessoas que colaboraram muitas vezes com poucos co-autores. Por fim, existe descontinuidade em alguns pontos, como

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

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 freqüência das arestas com esse peso.

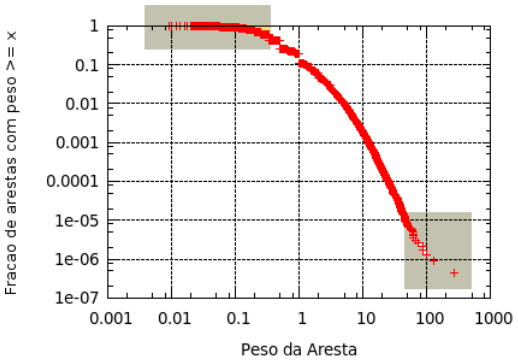

Figura 3. Distribuição dos pesos das arestas da rede de colaboração mundial.

Peso do Vértice

O peso do vértice corresponde ao número de publicações que o autor tem em colaboração com pelo menos um co-autor. Observando a figura 4 percebemos que a distribuição do peso dos vértices possui uma cauda pesada, ocorrendo valores muito distantes do peso médio, que é 3,9, variando de 0 a 529. [3] Outro dado interessante visto a partir do gráfico é que apenas 10% dos autores colaboraram em mais de 8 artigos. Logo, pode-se concluir que a grande maioria das pessoas colabora muito pouco e que poucas pessoas colaboram muito.

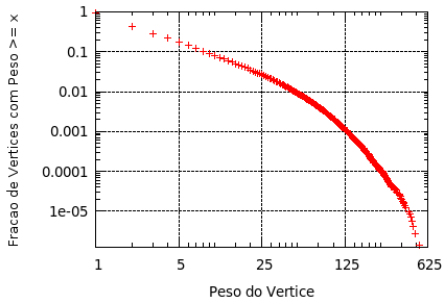

Figura 4. Distribuição dos pesos dos vértices da rede de colaboração mundial.

B. Análise da Rede Brasileira

A rede brasileira foi extraída da rede de colaboração mundial, considerando apenas os vértices e arestas dos pesquisadores que atuam no Brasil (subgrafo induzido). A *DBLP* não classifica os autores por nacionalidade nem por local de trabalho, então foi necessário desenvolvermos 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 (URL) de sua página pessoal. Então, de início, consideramos todos os autores com página pessoal contendo .br/ 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.

Para aumentar essa lista, acrescentamos os pesquisadores (presentes na *DBLP*) dos programas de pós-graduação em Ciência da Computação das universidades brasileiras disponíveis em [14] e os pesquisadores com bolsa de produtividade em pesquisa do CNPq, coletados a partir de [15].

Para identificar potenciais colaboradores de pessoas já identificadas como atuantes no Brasil, analisamos os vértices vizinhos ao conjunto dos brasileiros já selecionados e, manualmente, verificamos se o autor atuava no Brasil. Para cada autor verificado, quando este atuava no Brasil, adicionávamos ao conjunto de brasileiros. Assim, o conjunto de pessoas que atuam no Brasil ficou com exatamente 2.729 pessoas.

A rede brasileira avaliada contém 2.729 vértices e 6.953 arestas entre eles. Utilizaremos as mesmas métricas da seção III-A para caracterizar esta rede.

Grau

Observando a figura 5 percebemos que a distribuição do número de colaboradores que atuam no Brasil possui uma cauda pesada, ocorrendo valores muito distantes do grau médio que é de 5,1, variando de 0 a 101. Os indivíduos que atuam no Brasil têm em média menos colaboradores que a rede de colaboração mundial. Há um pequeno número de autores com muitas colaborações e uma grande quantidade de pessoas com poucos co-autores, por exemplo, no gráfico observamos 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, notamos 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.

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 6 percebemos que a distribuição do peso das

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

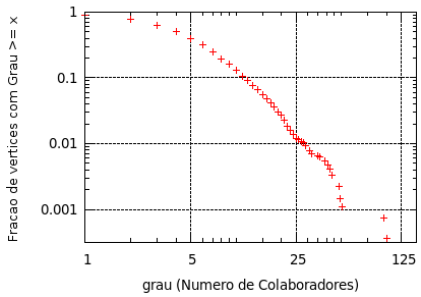

Figura 5. Distribuição do grau na rede de colaboração brasileira.

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.

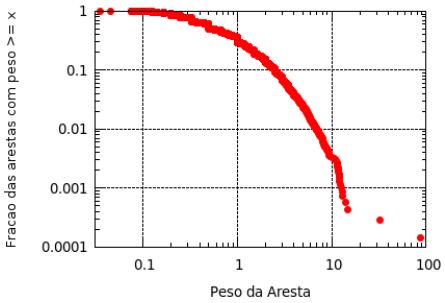

Figura 6. Distribuição dos pesos das arestas na rede de colaboração brasileira.

Peso dos Vértices

O peso médio de um vértice desta rede é 5,6, ou seja, o número médio de publicações em colaboração, um valor médio maior que na rede de colaboração mundial.

Observando a figura 7 percebemos que a distribuição do peso dos vértices possui uma cauda pesada, ocorrendo valores muito distantes do peso médio, variando de 0 a 123. Vértices com peso igual a zero correspondem a pessoas que possuem todas as suas publicações sozinhas, sem nenhum co-autor. 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.

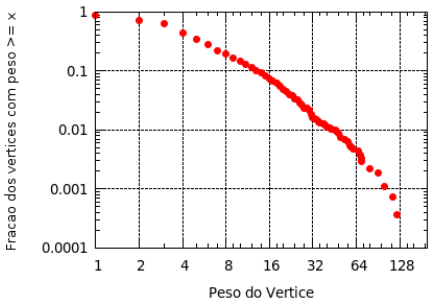

Figura 7. Distribuição dos pesos dos vertices na rede de colaboração brasileira.

Resumo das Métricas

A Tabela I apresenta os resultados numéricos da análise da estrutura topológica da rede. Dentre as métricas apresentadas, os valores de peso médio do vértice e grau médio da rede mundial e da rede brasileira chamam atenção. Constatamos que o peso médio do vértice é maior na rede brasileira, mas seu 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.

Tanto a rede de colaboração mundial quanto a rede de colaboração brasileira possui uma componente conexa gigante. Uma componente conexa gigante é o maior subgrafo conectado da rede, mais precisamente, é o maior conjunto de vértices, tal que todos os vértices conseguem alcançar os demais. Na rede de mundial constitui 79,8% dos vértices da rede, enquanto que na rede brasileira constitui 85,7%.

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 há quase 50% de chance de essa colaboração existir. Altos índices de clusterização são comuns em redes sociais. A tabela contém outras métricas que não foram discutidas por falta de espaço.

Outros trabalhos que apresentam resultados numéricos de análise da estrutura topológica de redes de colaboração em Ciência da Computação estão disponíveis em [10], [3] e [9].

IV. Métrica para Ranqueamento Baseada em Intensidade de Relacionamento

Apresentaremos nesta seção uma métrica para ranqueamento dos vértices baseada na intensidade de relacionamento. A intensidade dos relacionamentos está

Tabela I
Resumo das Métricas

	Mundial	Br
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 pub/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
Núm. de comp. conexas	77.493	297
Tam. médio das comp. conexas	9,3	9,2
Coef. 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

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

A métrica proposta utiliza a idéia de corte e de peso do vértice no corte de 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 outra ponta em $V - X$, onde V é o conjunto de todos os vértices pertencentes ao grafo G [16]. O peso do vértice no corte é dado pela soma dos pesos das arestas que definem o corte e incidem no vértice. Ou seja, dado um conjunto X de vértices, o peso de um vértice do conjunto X corresponde à soma dos pesos de todas as suas arestas que têm uma ponta neste vértice e a outra em algum vértice pertencente a $V - X$. Logo, os pesos das arestas formadas entre os vértices do conjunto X não contribuem no peso do vértice no corte. 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 vértice corte. A figura 8 será utilizada para exemplificar como medir o peso do vértice no corte.

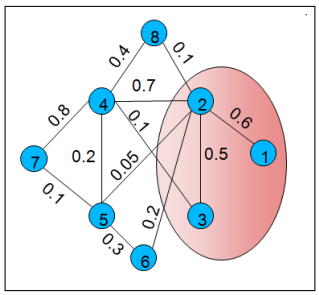

Figura 8. Exemplo de cálculo do peso do vértice no corte.

A figura 8 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.

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.

Intuitivamente, a métrica deve funcionar melhor quando |X| << |V|, ou seja, o número de elementos do subconjunto a ser ranqueado for bem menor que o número total de elementos. No entanto, essa análise foi deixada para trabalhos futuros, dado que nosso objetivo é ranquear um subconjunto dos vértices para o próprio subconjunto e não ranquear todos os vértices da rede.

Avaliação

Faremos o ranqueamento dos pesquisadores que atuam no Brasil utilizando quatro métricas objetivas, dentre elas a métrica proposta neste artigo. 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, comparamos 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. O termo avaliação subjetiva foi utilizado porque a avaliação não é baseada em uma métrica objetiva, por exemplo, uma equação. 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), nas quais 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 [17]. A relação de bolsistas de produtividade em pesquisa foi obtida em [15]. Todos estes indivíduos foram identificados manualmente na DBLP.

Na comparação dos resultados, utilizamos as métricas de precisão, cobertura e medida-F para verificar o grau de acerto de cada uma das quatro métricas de ran-

queamento com relação aos pesquisadores classificados de acordo com o CNPq. Precisão é a proporção de um conjunto de objetos recuperados que é realmente relevante. Cobertura é a proporção de objetos relevantes que foram recuperados [18] [19]. Medida-F é uma média harmônica que considera ao mesmo tempo precisão e cobertura.

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.

Fizemos 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, observamos 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 podemos calcular a precisão e cobertura e medir a eficácia da métrica em identificar os pesquisadores de nível 1A.

De posse dos ranqueamentos de cada métrica, consideramos listas contendo os n pesquisadores mais bem colocados em cada ranqueamento, onde n assumiu os valores de 20, 25, 30, 35 e 50, correspondendo a 0,7%, 0,9%, 1,1%, 1,3% e 1,8% dos pesquisadores que atuam no Brasil.

A tabela II apresenta a cobertura 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 cobertura 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 cobertura. 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.

Uma observação importante acerca da tabela II é que o peso do vértice e o número de publicações possuem o mesmo valor de precisão e cobertura para os tamanhos 20, 30 e 35 das listas de retornados. Isso se deve à semelhança das duas métricas, pois o peso do vértice corresponde ao número de publicações com ao menos uma colaboração. Pode-se concluir a partir desta observação, que os pesquisadores de nível 1A, em geral, publicam em colaboração com outros não 1A.

Tabela II
Precisão e cobertura dos ranqueamentos utilizando métricas diferentes para retornar pesquisadores com bolsa de produtividade de pesquisa 1A e variando o tamanho da lista.

Tamanho da Lista	Número de Publicações		Peso do Vértice		Peso do Vértice no Corte		Grau	
	Precisão	Cobertura	Precisão	Cobertura	Precisão	Cobertura	Precisão	Cobertura
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

Uma comparação direta entre as métricas pode ser vista na figura 9 que apresenta 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, acrescentamos os bolsistas de produtividade em pesquisa de nível 1B no conjunto de objetos relevantes e passamos a medir a eficácia da métrica ao recuperar os pesquisadores 1A e 1B, 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 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 corte, porém as métricas grau e número de publicações se mostram muito eficazes 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.

V. Conclusão

Neste artigo, apresentamos um estudo de algumas propriedades topológicas da rede de colaboração mundial e da rede de colaboração brasileira de autores de artigos científicos da área de Ciência da Computação. Em particular, caracterizamos a intensidades dos relacionamentos e dos vértices. Também desenvolvemos uma métrica para ranqueamento de vértices em redes de colaboração baseada na intensidade de relacionamento de um indivíduo com indivíduos fora de sua comunidade (conjunto de vértices).

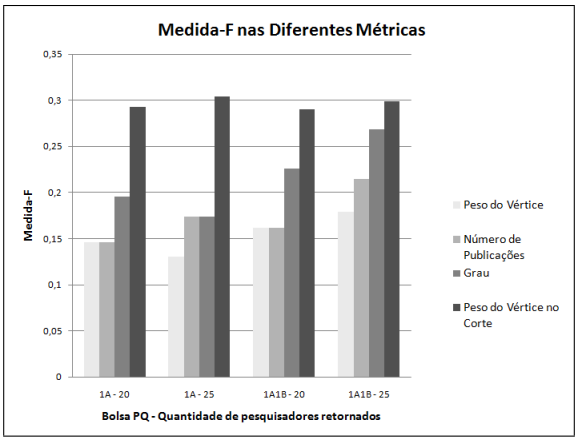

Figura 9. Medida-F aplicada nas quatro métricas de ranqueamento ao retornar pesquisadores de nível 1A e 1B.

Para avaliar a eficácia da métrica, comparamos os ranqueamentos de pesquisadores da área de Ciência da Computação do Brasil feitos a partir de métricas objetivas com a classificação subjetiva feita pelo CNPq através da bolsa de produtividade em pesquisa. Nossos resultados indicam que a métrica proposta nesse artigo é eficaz na identificação de pesquisadores mais influentes no Brasil.

Para trabalhos futuros, nós planejamos 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 idéia surgiu a partir da intuição de que colaborações antigas têm menor importância na intensidade do relacionamento entre dois indivíduos.

VI. Agradecimentos

Este trabalho recebeu o apoio financeiro da CAPES (bolsa de mestrado) e da FAPERJ através do Programa Jovem Cientista do Nosso Estado.

Referências

[1] M. E. J. Newman, "The structure of scientific collaboration networks," in *Proc. Natl. Acad. Sci. USA*, vol. 98, Jan. 2001, pp. 404–409.

[2] C. Wagner and L. Leydesdorff, "Network structure, self-organization, and the growth of international collaboration in science," *Research Policy*, vol. 34, no. 10, pp. 1608–1618, December 2005.

[3] M. E. J. Newman, *Who Is the Best Connected Scientist? A Study of Scientific Coauthorship Networks*, ser. Lecture Notes in Physics, Aug. 2004, vol. 650/2004, pp. 337–370.

[4] V. S. de Andrade Menezes, R. T. da Silva, M. F. de Souza, J. Oliveira, C. E. R. de Mello, J. M. de Souza, and G. Zimbrão, "Mining and analyzing organizational social networks using minimum spanning tree," in *OTM Confederated International Workshops and Posters on On the Move to Meaningful Internet Systems*, 2008.

[5] C. Barcellos and F. I. Bastos, "Redes sociais e difusão da aids no brasil," *Organización Panamericana de la Salud*, 1996.

[6] M. C. P. I. Hayashi, C. R. M. Hayashi, and M. Y. de Lima, "Análise de redes de co-autoria na produção científica em educação especial," *Liinc em Revista*, vol. 4, no. 1, pp. 84–103, Mar. 2008.

[7] J. Grossman and P. Ion, *The Erdös Number Project*. [Online]. Available: http://www.oakland.edu/enp/

[8] J. Huang, Z. Zhuang, J. Li, and C. L. Giles, "Collaboration over time: characterizing and modeling network evolution," in *WSDM '08: International conference on Web search and web data mining*. New York, NY, USA: ACM, 2008, pp. 107–116.

[9] M. E. J. Newman, "Coauthorship networks and patterns of scientific collaboration," in *Proc. Natl. Acad. Sci. USA*, vol. 101, no. Suppl 1, Apr. 2004, pp. 5200–5205.

[10] G. V. Menezes, N. Ziviani, A. H. F. Laender, and V. A. F. Almeida, "A geographical analysis of knowledge production in computer science," in *18th international conference on World wide web*, 2009, pp. 1041–1050.

[11] *The DBLP Computer Science Bibliography*. [Online]. Available: http://www.informatik.uni-trier.de/ ley/db/

[12] M. Ley, "Dblp — some lessons learned," *VLDB '09*, 2009.

[13] J. P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. L. Barabasi, "Structure and tie strengths in mobile communication networks," *National Academy of Sciences*, vol. 104, no. 18, pp. 7332–7336, Oct 2006.

[14] *Docentes da pós-graduação e bolsistas de produtividade em pesquisa*. [Online]. Available: http://ged.capes.gov.br/AgProd/silverstream/pages/ pgRelBolsistasProdPesqResultado.html

[15] CNPq, *Bolsas de Produtividade em Pesquisa - Bolsas em curso*. [Online]. Available: http://www.cnpq.br

[16] J. L. Szwarcfiter, *Grafos e algoritmos computacionais*. Campus, 1986.

[17] *CNPq - Instrumentos Normativos - Produtividade em Pesquisa (PQ)*. [Online]. Available: http://universia.com.br/materia/materia.jsp?id=7585

[18] N. J. Belkin and W. B. Croft, "Information filtering and information retrieval: two sides of the same coin?" *Commun. ACM*, vol. 35, no. 12, pp. 29–38, 1992.

[19] M. Zhu, *Recall, Precision and Average Precision*, 2004.