UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

GISELI RABELLO LOPES

Avaliação e Recomendação de Colaborações em Redes Sociais Acadêmicas

Tese apresentada como requisito parcial
para a obtenção do grau de
Doutor em Ciência da Computação

Prof. Dr. José Palazzo Moreira de Oliveira
Orientador

Prof[a]. Dr[a]. Mirella M. Moro
Coorientadora

Porto Alegre, maio de 2012

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Lopes, Giseli Rabello

Avaliação e Recomendação de Colaborações em Redes Sociais Acadêmicas / Giseli Rabello Lopes. – Porto Alegre: PPGC da UFRGS, 2012.

129 f.: il.

Tese (doutorado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2012. Orientador: José Palazzo Moreira de Oliveira; Co-orientadora: Mirella M. Moro.

1. Redes sociais. 2. Avaliação de qualidade. 3. Sistemas de recomendação. I. Oliveira, José Palazzo Moreira de. II. Moro, Mirella M.. III. Título.

"Descobri como é bom chegar quando se tem paciência. E para se chegar, onde quer que seja, aprendi que não é preciso dominar a força, mas a razão. É preciso, antes de mais nada, querer."

— AMYR KLINK

AGRADECIMENTOS

Meus sinceros agradecimentos a todos que me incentivaram e contribuíram de alguma forma para o desenvolvimento desta tese. Em especial, gostaria de agradecer:

Ao meu orientador, Prof. José Palazzo Moreira de Oliveira, por ter me conduzido a um amadurecimento em minha vida acadêmica, através de seus conselhos e oportunidades concedidos, primeiramente durante meu mestrado e agora como doutoranda. Agradeço imensamente por todas as discussões, sugestões e direções apontadas, que foram de extrema importância durante o andamento de meu doutorado. Sua ampla visão, experiência, conhecimento, entusiasmo, dedicação e competência foram essenciais para orientar-me, com sabedoria, ao longo de mais esta jornada.

À minha coorientadora, Profª. Mirella M. Moro, que muito me motivou através de seu grande exemplo de amor, entusiasmo e dedicação à academia. Agradeço por ter aceitado ser minha coorientadora e por toda a atenção que me foi dispensada desde então. Seus comentários, sugestões, discussões e revisões, sempre muito relevantes, contribuíram imensamente na condução do trabalho desenvolvido.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo imprescindível suporte financeiro que permitiu minha dedicação ao doutorado.

Ao Instituto de Informática da UFRGS, por toda a infraestrutura disponibilizada e a seus profissionais; aos professores do PPGC pelos ensinamentos e experiências transmitidos e aos funcionários sempre solícitos e prestativos.

Aos professores Roberto da Silva e Leandro Krug Wives, pelos conhecimentos e ideias compartilhados, e pelas importantes sugestões e contribuições em coautorias de artigos relativos a esta tese. À Profª. Viviane P. Moreira pelas importantes sugestões de alguns métodos para avaliação dos experimentos.

Aos professores Marcelo Soares Pimenta, Mirella M. Moro, Renata de Matos Galante e Stanley Loh, que participaram da minha banca de defesa de proposta de tese, pelas importantes análises e sugestões que incentivaram a continuidade e o aprimoramento do trabalho proposto.

Aos membros da banca de defesa desta tese: Prof. Dr. Eliseo Berni Reategui (PPGIE-UFRGS), Prof. Dr. Nivio Ziviani (UFMG) e Profª. Drª. Renata de Matos Galante (INF-UFRGS), por terem aceitado o convite, pelas sugestões que contribuíram no aprimoramento do texto final e pelo incentivo para o seguimento da pesquisa realizada.

Ao InWeb (Instituto Nacional de Ciência e Tecnologia para Web), pela oportunidade de poder fazer parte desta equipe, participar de discussões e reuniões e desenvolver um trabalho no contexto deste projeto. Através dessa parceria, agradeço a oportunidade de estada pelo período de um mês junto ao grupo do Laboratório de Banco de Dados (LBD) da Universidade Federal de Minas Gerais (UFMG), sob a orientação da Profª. Mirella M.

Moro, tendo sido de extrema importância para o estabelecimento de novas possibilidades de trabalho conjunto e para o aprimoramento das definições de algumas experimentações.

Ao aluno Eduardo M. Barbosa, por ter aceitado o desafio de implementar uma ferramenta de visualização de redes acadêmicas, partindo de definições desta tese.

A todos os meus amigos que acompanharam, de perto ou de longe, a realização deste trabalho e que torceram para que o mesmo fosse concluído com êxito. A todos os colegas do Grupo de Sistemas de Informação da UFRGS que tive a oportunidade de conhecer neste período, pelo apoio e incentivo proporcionados. Em especial, aos amigos e colegas de doutorado: Ana Marilza Pernas, Daniel Lichtnow, Eduardo Borges, Isabela Gasparini e Leila Weitzel, com os quais tive o privilégio de conviver mais diretamente e que propiciaram um ótimo ambiente de integração e troca de experiências; agradeço por partilharem esses anos de estudos e expectativas comigo. Agradecimentos muito especiais, aos queridos amigos: Eduardo Borges, Euler Taveira, Marcos Nunes e Otavio Acosta, que foram companheiros importantes pelo apoio, incentivo, amizade e por todos os momentos de seriedade e descontração partilhados no decorrer deste período de meu doutoramento.

Aos meus pais, Valdir e Elza, que desde o princípio foram grandes incentivadores para que eu me dedicasse aos estudos. Agradeço por todo amor, carinho, confiança e apoio recebidos os quais só comprovam que posso sempre contar com eles. Levarei para toda a minha vida suas importantes lições e ensinamentos. Além disso, agradeço também à minha mãe, Prof[a]. Elza Lopes, pelas correções ortográficas e gramaticais efetuadas nesta versão final.

À minha irmã, Cyntia, pelo carinho e torcida mesmo à distância.

Ao meu noivo, Daniel da Costa Mendes, por todo amor, carinho, amizade e companheirismo a mim dirigidos. Agradeço pela confiança que sempre depositou em mim, por apoiar minhas decisões, entender minhas ausências, incentivar minhas aspirações e ser conforto e segurança mesmo nos momentos adversos.

Por fim, agradeço imensamente a Deus por ter me guiado e conduzido durante mais esta importante etapa de minha vida.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

ACM	*Association for Computing Machinery*
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CLEF	*Cross Language Evaluation Forum*
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
DBLP	*Digital Bibliography & Library Project*
IDF	*Inverse Document Frequency*
IF	*Impact Factor*
ISI	*Institute for Scientific Information*
JCR	*Journal Citation Reports*
MAP	*Mean Average Precision*
SN	*Social Network*
SNA	*Social Network Analysis*
TF	*Term Frequency*
TREC	*Text REtrieval Conference*
VSM	*Vector Space Model*
XML	*eXtensible Markup Language*

GLOSSÁRIO DE TERMOS

Para padronização da terminologia utilizada no texto desta tese, em relação aos termos abaixo, foram adotadas as seguintes definições[1]:

Indicador Modelo e critérios de decisão, a fim de prover uma estimativa ou avaliação de um conceito calculável a partir de uma métrica.

Métrica A própria medida (métrica direta) ou um método de cálculo (métrica indireta) e a escala de medição (consideração no espaço métrico). Uma métrica determina padrões de medição pelos quais um determinado Indicador pode ser avaliado.

Função Algoritmo ou fórmula desenvolvida para combinar duas ou mais métricas (em um nível mais alto, pode ser vista como uma agregação de indicadores).

[1] Não existe um consenso em relação à terminologia (nem mesmo na ISO). Alguns termos são originários da matemática. Existem trabalhos sendo desenvolvidos na tentativa de padronizar as definições, incluindo o desenvolvimento de ontologias para definir estes termos (muitos relativos à qualidade de software), como o caso de (OLSINA; ANGELES MARTÍN, 2004) que serviu de base principal para as definições utilizadas nesta tese.

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

No contexto acadêmico o trabalho de pesquisa científica, nas áreas tecnológicas, é efetuado através de colaborações e cooperações entre diferentes pesquisadores e grupos de pesquisa. Existem pesquisadores atuando nos mais variados assuntos e nas mais diversas subáreas de pesquisa. Para analisar e expandir tais colaborações, muitas vezes, é necessário avaliar o nível de cooperação dos atuais parceiros, bem como identificar novos parceiros para conduzir trabalhos conjuntos. Tal avaliação e identificação não são tarefas triviais. Dessa forma, abordagens para avaliação e recomendação de colaborações são de grande valia para o aperfeiçoamento da cooperação e consequente melhoria da qualidade da pesquisa.

Em relação à análise de colaborações, a demanda por critérios de avaliação de qualidade e por métodos de avaliação associados está aumentando e tem sido foco de muitos estudos na última década. Esse crescimento surge devido à busca por excelência acadêmica e para o apoio à tomada de decisões por parte de agências de financiamento para a alocação de recursos. Nesse contexto, há uma tendência a empregar técnicas bibliométricas, especialmente métodos estatísticos aplicados a citações. Com tanto material sendo pesquisado e publicado, resolveu-se explorar outra faceta para definição de indicadores de qualidade no contexto acadêmico visando a obtenção de resultados complementares e que garantam, através de sua validação experimental, uma melhor geração de indicadores. Desse modo, nesta tese, utiliza-se a tendência atual de estudos em análises de redes sociais, definindo métricas sociais específicas para definição de tais indicadores. Neste trabalho, é apresentada uma função para avaliação de qualidade de grupos de pesquisa com base nas colaborações internas entre seus pesquisadores membros. Estas colaborações são avaliadas através de análises em redes sociais bibliográficas acadêmicas baseadas em métricas de interação social.

Com relação à identificação ou recomendação de colaborações, esta tese apresenta uma abordagem que considera tanto a parte de conteúdo quanto a de estrutura de uma rede. Especificamente, o conteúdo envolve a correlação entre os pesquisadores por áreas de pesquisa, enquanto a estrutura inclui a análise da existência de relacionamentos prévios entre os pesquisadores. Grande parte das abordagens que efetuam a recomendação de colaborações foca em recomendar especialistas em uma determinada área ou informação. Essas não consideram a área de atuação do usuário alvo da recomendação, como no caso da abordagem apresentada nesta tese. Além disso, neste trabalho, a obtenção de informações sobre os relacionamentos entre usuários, para construção de uma rede social acadêmica, é feita de forma implícita, em dados sobre publicações obtidos de bibliotecas digitais. Utilizando tais dados, também é possível explorar aspectos temporais para ponderação desses relacionamentos, utilizando-os para fins de recomendação de colaborações. Não foram encontrados trabalhos prévios nesse sentido. A presente abordagem

inclui a recomendação não só de novas colaborações, como também, a recomendação de intensificação de colaborações já existentes, o que não é considerado por outros trabalhos relacionados. Dessa forma, pode-se dizer que os objetivos de recomendação da presente abordagem são mais amplos.

Após propor novas técnicas para avaliação e identificação de parcerias, esta tese as valida através de uma avaliação experimental. Especificamente, experimentos com dados reais sobre as relações de coautoria entre pesquisadores pertencentes a diferentes grupos de pesquisa são apresentados para avaliação e demonstração da validade e da aplicabilidade das diferentes proposições desta tese referentes à avaliação de qualidade e recomendação de colaborações.

Palavras-chave: Redes sociais, avaliação de qualidade, sistemas de recomendação.

Evaluation and Recommendation of Collaborations on Academic Social Networks

ABSTRACT

In technological fields, scientific research is performed through collaboration and cooperation of different researchers and research groups. In order to analyze and expand such collaborations, it is necessary to evaluate the level of cooperation between current partners as well as to identify new partners. Such an analysis and identification are not trivial tasks. Thus, approaches to evaluating and recommending collaborations are valuable to improve cooperation and, hence, improve research quality.

Regarding the collaborations evaluation, the demand for quality assessment criteria and associated evaluation methods is increasing. Indeed, such evaluations have been the focus of many studies in the last decade. This growth arises from the pursuit of academic excellence and decision making of funding agencies. In this context, the trend is to employ bibliometric techniques, especially citation statistics. With so much material being researched and published, another facet for defining quality indicators is explored. Our goal is to obtain additional results that ensure, through its experimental validation, a better indicators generation. In this thesis, the current trend of studies in social network analysis is applied in the definition of such indicators. Specifically, we introduce a function for quality assessment of research groups based on internal collaborations among their member researchers. These collaborations are evaluated through analysis on bibliometric academic social networks based on metrics of social interaction.

Regarding the collaborations recommendation, this thesis presents an approach that considers both the content and structure of research networks. The content involves the correlation among researchers by research areas whereas the structure includes the analysis of existing relationships among researchers. Most of the approaches that perform the collaborations recommendation focus on recommending experts in a certain area or information. They do not consider the working area of the recommendation target user, as we do in this thesis. Moreover, here, the information about the researchers' relationships, employed for building an academic social network, is implicitly obtained through publications data available in digital libraries. Moreover, we expand previous analysis by considering temporal aspects to determine the relationships weights (which may be used to collaborations recommendation purposes). There were no previous studies in this direction. Our approach includes not only the recommendation of new collaborations, but also the recommendation of the collaborations intensification, which is not considered by other related work.

After proposing new techniques for evaluating and identifying research collaborators, this thesis validates it through an experimental evaluation. Specifically, we evaluate and demonstrate the applicability of our techniques considering real datasets on the co-author relationships among researchers from different research groups.

Keywords: Social networks, quality assessment, recommender systems.

1 INTRODUÇÃO

A Web 2.0 é a segunda geração de aplicações integrando comunidades e serviços caracterizada pelo provimento de técnicas para publicação pessoal, compartilhamento, colaboração e organização de informações na *World Wide Web*. A Web 2.0 tem significativas implicações sociais, que levam aos processos de trabalho coletivo e colaborativo, à troca afetiva, produção e circulação de informação, e construção de conhecimento social suportados pela tecnologia da informação. Nessa perspectiva, não somente os aspectos tecnológicos e de conteúdo, mas também as interações sociais e seus aspectos relacionais devem ser levados em consideração. Nesse contexto, as comunidades, serviços, e aplicações baseadas na Web emergiram, incluindo as redes sociais online, que são aplicações Web muito interessantes. Exemplos de tais redes incluem *LinkedIn*[1], *Facebook*[2], *Orkut*[3], *Google+*[4], entre outras, e cada uma dessas conecta milhões de usuários.

O interesse de acadêmicos e do público em geral sobre as redes sociais tem crescido rapidamente. Especificamente, o crescente interesse na pesquisa sobre Redes Sociais (*Social Networks* - SN) foi incentivado pela popularização das redes sociais online. Um tópico relacionado é a Análise de Redes Sociais (*Social Networks Analysis* - SNA), que supõe que a interação entre as unidades é o ponto central para a avaliação e análise de colaboração social. A perspectiva de rede social inclui teorias, modelos e aplicações que são expressos em termos de conceitos relacionais. Além disso, a medição rigorosa da relação definida pelas ligações entre as partes é um componente fundamental para inferir propriedades da rede estudada.

Alguns conceitos fundamentais utilizados em SNA incluem atores e vínculos relacionais (KNOKE; YANG, 2007; WASSERMAN; FAUST, 1994). Atores são entidades sociais que têm suas ligações sociais modeladas por uma rede social. Eles são ligados a outros por vínculos relacionais. O intervalo e o tipo desses vínculos podem ser bastante extensos. Portanto, uma característica significante e crítica de uma rede social é a presença de informação relacional. O crescente interesse mencionado anteriormente e o uso da análise de redes sociais têm provido um consenso sobre os princípios predominantes das redes que distinguem análises de redes sociais de análises desenvolvidas para outros tipos de redes.

Desenvolver métodos no contexto de SNA é significativamente importante, porque a análise de uma unidade particular em uma rede social não envolve um único indivíduo, mas uma entidade consistindo em coleções de indivíduos e ligações entre eles. Mais ainda, SNA é um esforço inerentemente interdisciplinar: seus conceitos são desenvolvidos

[1]LinkedIn: http://www.linkedin.com/
[2]Facebook: http://www.facebook.com/
[3]Orkut: http://www.orkut.com
[4]Google+: https://plus.google.com/

em teoria social, pesquisa empírica e matemática formal e estatística (WASSERMAN; FAUST, 1994). Além disso, os pioneiros em SNA vieram da sociologia, psicologia social e antropologia, e o primeiro uso do termo "rede social" foi atribuído a Barnes (1954).

Os métodos de SNA provêm declarações formais sobre processos e propriedades sociais. Além disso, esses conceitos podem ser definidos precisamente e consistentemente. Uma vez definidos, os mesmos podem fornecer novas perspectivas sobre o mundo social (FREEMAN, 1984). Trabalhos recentes empregam SNA para entender interconexões, evoluções e comportamentos de comunidades de pesquisa como um todo (DING, 2011; MENEZES et al., 2009; WANG et al., 2010).

Em um contexto mais específico e diferente da tradicional rede social de amigos, uma rede social *acadêmica* pode representar colaborações científicas. Nessa, atores representam pesquisadores e vínculos relacionais representam os relacionamentos (colaborações) entre esses. O relacionamento pode ser dado por qualquer tipo de interação entre dois pesquisadores. Por exemplo, a presença de pelo menos um artigo em coautoria entre dois pesquisadores pode determinar um vínculo relacional entre eles. Para identificar tais relações, podem ser usadas diferentes fontes de dados disponíveis fisicamente ou na Web, como por exemplo DBLP[5] (*Digital Bibliography & Library Project*), Google Scholar[6], CiteSeer[7], BDBComp[8], entre outras.

Além disso, na comunidade científica, é muito comum criar métricas para praticamente tudo que possa ser mensurado. Existem eventos e publicações especializados em métricas de diferentes perspectivas, tais como a conferência *International Conference on Scientometrics and Informetrics* e os periódicos *Scientometrics* e *Informetrics*. Na verdade, a perseguição por excelência em áreas de pesquisa e a competição por recursos financeiros têm motivado estudos da avaliação de qualidade da própria pesquisa, principalmente utilizando métricas bibliométricas (EGGHE, 2006, 2010; HABIBZADEH; YADOLLAHIE, 2008; HIRSCH, 2005; NICOLAISEN; FRANDSEN, 2008; REN; TAYLOR, 2007).

Essas métricas podem ser aplicadas não somente para determinar qualidade em pesquisa mas também: para ajudar na tomada de decisão das agências de fomento, para avaliar qualidade de características relacionadas à pesquisa para alocação orçamental e até para alocação dos recursos humanos. A grande pressão em tal cenário requer a definição de indicadores associados à avaliação de qualidade que possam ser objetivamente definidos (através de métricas para sua estimação) e, preferencialmente, facilmente reproduzidos. Essa última característica é importante porque o resultado de qualquer estratégia de ranking pode ser questionado depois de sua publicação. Por isso, é desejável permitir que os resultados possam ser reproduzidos e verificados por qualquer pessoa.

Especificamente sobre a avaliação da qualidade de grupos de pesquisadores, métricas desenvolvidas para esse fim vêm sendo empregadas para definição de diferentes rankings, não só do grupo diretamente, tais como: rankings de periódicos e conferências com base na qualidade de seu corpo editorial e comitê de programa, rankings de universidades com base na qualidade de seus pesquisadores e membros do corpo docente, e rankings de propostas de projetos de pesquisa com base na qualidade de seus pesquisadores proponentes. Esse é um ponto muito sensível, pois o desenvolvimento ou não de um grupo de pesquisa pode ser consequência de tal avaliação. Nesse contexto, há uma tendência

[5]DBLP: http://www.informatik.uni-trier.de/~ley/db
[6]Google Scholar: http://scholar.google.com
[7]Citeseer: http://citeseer.ist.psu.edu/index
[8]BDBComp: http://www.lbd.dcc.ufmg.br/bdbcomp/

de empregar técnicas bibliométricas, especialmente estatísticas de citações (MOLINARI; MOLINARI, 2008; REN; TAYLOR, 2007; SILVA et al., 2010; YAN; LEE, 2007). No entanto, uma aplicação ingênua de métricas de bibliometria pode facilmente levar a uma classificação injusta.

Com tudo que vem sendo pesquisado e publicado em bibliometria, resolveu-se explorar outra faceta para definição de indicadores de qualidade no contexto acadêmico. Nesta tese, resolveu-se aproveitar a tendência atual de estudos em análises de redes sociais e aplicá-la na definição de tais indicadores. É importante notar que unindo as poderosas análises providas por SNA e os dados disponibilizados sobre comunidades de pesquisa, podem ser definidas métricas para avaliar a forma como os grupos de pesquisa colaboram. Mais ainda, SNA permite analisar colaborações entre pesquisadores bem como quantificar comportamentos de interação científica.

O trabalho desta tese está inserido no contexto de avaliação de qualidade na academia. Ao contrário dos trabalhos relacionados anteriormente (em bibliometria), o presente trabalho não considera estatística de citações. São explorados outros indicadores que podem ser usados para avaliar a qualidade de grupos de pesquisadores (para fins de experimentação cada grupo é constituído por pesquisadores de um mesmo programa de Pós-graduação) associados à análise das colaborações internas. Portanto, diferentemente dos trabalhos prévios, são introduzidas diferentes métricas de análise de redes sociais para esse fim. Trabalhos relacionados em redes sociais utilizam análises com propósitos comparativos e para entender o comportamento das interações, mas eles não objetivam inferir qualidade ou construir um ranking de grupos como é o caso desta tese. Cabe destacar que este trabalho foi aplicado em um estudo de caso para avaliar programas de Pós-graduação com base na colaboração interna de seus grupos de pesquisadores. Entretanto, o mesmo pode ser aplicado em outros contextos onde a interação entre partes é um dos pontos a ser analisado na definição de qualidade.

Um dos únicos trabalhos encontrados na literatura que faz uso de análises de redes sociais com objetivos de ranqueamento acadêmico é o de Freire e Figueiredo (2010). Porém, seus autores exploram outro indicador de qualidade, o qual é definido em relação às colaborações externas entre pesquisadores. Especificamente, é proposta uma métrica em relação ao número de ligações quando é efetuado um corte em uma rede social de coautoria, relativo às ligações externas ao grupo sendo avaliado. Seu objetivo principal foi ranquear indivíduos.

Além disso, em geral, pesquisa científica é comumente desenvolvida através de colaboração envolvendo indivíduos de diferentes especialidades e perfis. Essa cultura de publicação é uma consequência da necessidade de grupos de pesquisa. Um dos mais conhecidos exemplos é o ATLAS de pesquisa da física da partícula (*particle physics*) do CERN em Genève[9], um esforço de pesquisa apoiado por 38 países, com um total de 3.000 físicos e 1.000 estudantes advindos de mais de 174 universidades e laboratórios. O ATLAS vem empreendendo um dos maiores esforços colaborativos já realizados nas ciências físicas. Nesse ambiente cooperativo, a multiplicidade de coautores é natural. Entretanto, em outras áreas, tais como Ciências Sociais, a autoria é mais individualizada, pois a criação de conhecimento é uma tarefa individual. Esta tese focaliza o primeiro cenário no qual colaboração é vital, incluindo uma ampla gama de áreas de pesquisa de Ciência da Computação, Engenharia e Física à Biologia e Medicina.

Em tais cenários orientados a grupo, analisar o nível de cooperação dos atuais parceiros bem como identificar novos parceiros para conduzir trabalhos em conjunto podem ser

[9] ATLAS Experiment: http://atlas.ch

necessários para diferentes propósitos e definitivamente não são tarefas triviais. Por exemplo, identificar novos parceiros de pesquisa é necessário quando se procura colaboradores para aplicações de fundos de pesquisa, move-se ou se é inserido em uma nova instituição, e para melhorar o status de colaboração de um grupo de pessoas. Dessa forma, a definição de abordagens para recomendação de colaborações é de grande valia. Especificamente, esse problema pode ser mapeado para uma *função de recomendação* que dado um usuário alvo u e um conjunto de dados de pesquisadores R, retorna uma lista ranqueada de indivíduos $r_i \in R$ tal que uma colaboração entre u e r_i é positivamente indicada.

Sistemas de recomendação surgiram visando a redução dos problemas associados ao fenômeno da sobrecarga de informação, procurando minimizar o tempo gasto para acessar a informação relevante. Sistemas de recomendação envolvem personalização da informação. A personalização está relacionada com o modo pelo qual a informação e serviços podem ser ajustados às necessidades específicas de um usuário ou comunidade (SMEATON; CALLAN, 2005). Tradicionalmente, sistemas de recomendação são estudados em três diferentes perspectivas, de acordo com as metodologias utilizadas para realizar a recomendação: (i) *filtragem baseada em conteúdo*, que recomenda itens classificados de acordo com um perfil do usuário e suas escolhas prévias; (ii) *filtragem colaborativa*, que lida com similaridades entre interesses dos usuários; e (iii) *filtragem híbrida*, que combina as duas anteriores para obter vantagens de seus benefícios.

Entretanto, sistemas de recomendação estão inseridos em um contexto social, uma vez que as recomendações são entregues a um usuário ou a uma comunidade de usuários. Perugini, Gonçalves e Fox (2004) enfatizam que a recomendação tem um inerente elemento social e, em última análise, destina-se a conectar pessoas, quer diretamente como resultado da modelagem explícita do usuário, quer indiretamente através da descoberta de relacionamentos implícitos nos dados existentes. Muito embora métricas e técnicas de análise em redes sociais venham sendo aplicadas para o entendimento de como um grupo de pessoas interage, apenas recentemente, seu potencial para indicar novas conexões tem sido enfatizado.

Dada a complexidade de sistemas de recomendação (que definem mecanismos de entrada e saída para uma função de recomendação), muita pesquisa tem sido realizada para otimizar diferentes aspectos tais como a geração e manutenção do perfil do usuário (BURKE, 2002; GEYER et al., 2008; MONTANER; LÓPEZ; ROSA, 2003), a função de recomendação (ou técnica de filtragem) (ADOMAVICIUS; TUZHILIN, 2005; BALABANOVIC; SHOHAM, 1997) e as conexões dos usuários (PERUGINI; GONÇALVES; FOX, 2004). Além disso, diferentes técnicas empregam o contexto social para melhorar a recomendação de itens como filmes (GOLBECK; HENDLER, 2006), artigos (HWANG; WEI; LIAO, 2010; WENG; CHANG, 2008; ZAIANE; CHEN; GOEBEL, 2007) e media social (GUY et al., 2010). Tais abordagens consideram principalmente as conexões estabelecidas dentro das redes sociais, frequentemente considerando aspectos somente topológicos, estruturais da rede. Por outro lado, muitas informações ricas podem ser extraídas de redes sociais e das conexões modeladas nas mesmas, pelo uso de métricas especificamente desenvolvidas, por exemplo.

Especificamente no cenário acadêmico, é muito importante entender que a adaptação de um sistema de recomendação para tal cenário não é uma tarefa trivial. Por exemplo, para sugerir novos amigos em uma rede social qualquer, um número de amigos em comum pode ser preponderante para tanto. Porém, em um contexto acadêmico, outros indicadores devem ser analisados. Deve-se considerar, por exemplo, as conexões em trabalhos prévios, artigos em coautoria, áreas de pesquisa em comum e muitos outros. Além

disso, o tipo de recomendação pode ser diferenciado. O trabalho desta tese visa considerar essa visão e colocar em ação a recomendação não apenas de novas colaborações, mas também da intensificação de colaborações já existentes. No geral, pode-se dizer que o objetivo final da abordagem desta tese é mais amplo do que o dos trabalhos anteriores, uma vez que estimula novas colaborações, bem como a intensificação de colaborações previamente existentes que são extremamente relevantes no contexto acadêmico.

Muitas abordagens em recomendação que utilizam redes sociais consideram alguns aspectos estruturais da rede para gerar recomendações. Métricas originárias em teoria de grafos são empregadas para estabelecer escores que estimam a proximidade entre os atores (modelados por uma Rede Social) e até mesmo para fazer a predição de novas ligações (LIBEN-NOWELL; KLEINBERG, 2007; NEWMAN, 2003; QUERCIA; CAPRA, 2009). Um estudo detalhado sobre o uso de diferentes métodos para predição de ligações em redes sociais é apresentado em (LIBEN-NOWELL; KLEINBERG, 2007). Entretanto, esse trabalho não explora métodos de estabelecimento de pesos para as ligações, nem o impacto desses nos resultados das funções de escore.

Existe uma infinidade de sistemas de recomendação que geralmente trabalham em domínios distintos. O trabalho desta tese tem seu foco na recomendação de colaboradores dentro do contexto acadêmico. Abordagens anteriores mais relacionadas incluem os trabalhos sobre recomendação de especialistas (*experts*) (KAUTZ; SELMAN; SHAH, 1997; MCDONALD, 2003). Essas abordagens não consideram a área de atuação do usuário alvo da recomendação e do usuário recomendado como é o caso da presente abordagem, mas somente do usuário recomendado. Existem também as abordagens que efetuam a recomendação de colaborações, com base na correlação de usos e interesses. Essas, geralmente, fazem uso de avaliações explícitas por parte dos usuários desse tipo de sistema, o que não ocorre na abordagem desta tese que visa a trabalhar com a obtenção de informações sobre os relacionamentos entre usuários de forma implícita baseada em dados, sobre publicações, obtidos de bibliotecas digitais.

Em redes sociais acadêmicas (redes de coautoria), a rede pode não ter a definição dos pesos das ligações (KAUTZ; SELMAN; SHAH, 1997; LIBEN-NOWELL; KLEINBERG, 2007; MCDONALD, 2003). Por outro lado, para uma rede com a definição dos pesos das ligações, o coeficiente de *Jaccard* é usualmente aplicado (ALEMAN-MEZA et al., 2006; HWANG; WEI; LIAO, 2010). No trabalho desta tese, vai-se um passo adiante, e pela primeira vez, é explorada a influência dos aspectos temporais nos pesos das ligações que são utilizados para determinar os resultados das recomendações. Recentes propostas têm abordado a importância de considerar aspectos temporais, na recomendação de itens dentro de um contexto social, para recomendação de itens (HWANG; WEI; LIAO, 2010; XIANG et al., 2010) (o primeiro é temporal no nível do perfil focado em tarefa e o segundo é temporal em relação às preferências do usuário) e para efetuar análises de redes sociais, não com propósitos de recomendação (TANG et al., 2009). Dessa forma, por não terem sido encontrados outros trabalhos correlatos, pode-se dizer que o trabalho desta tese é um dos primeiros a estudar a influência de aspectos temporais (através da ponderação das relações na SN considerando esse aspecto) na recomendação de colaboradores para Redes Sociais Acadêmicas.

1.1 Objetivos e Contribuições

Em um mundo em crescente “conexão”, esta tese tem o objetivo de investigar os desafios de pesquisa em redes sociais acadêmicas e propor soluções para os mesmos.

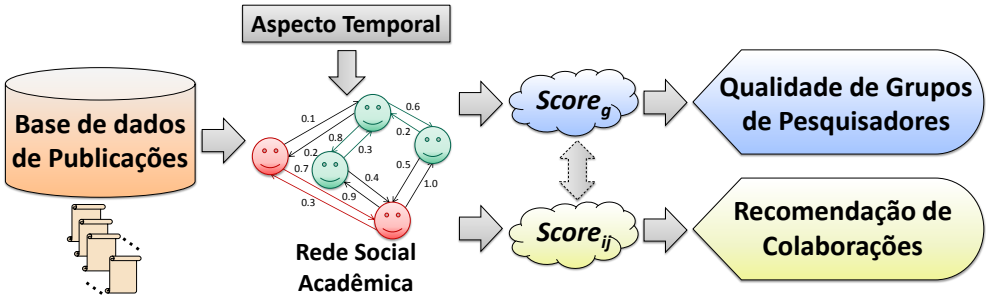


Figura 1.1: Visão geral da tese.

Especificamente, são explorados os desafios em:

- Avaliação de colaborações e determinação de indicadores de qualidade relacionados, para grupos de pesquisadores;
- Recomendação de colaborações científicas.

O propósito é determinar os indicadores importantes de serem analisados e refinados em ambos os contextos e propor diferentes métricas para sua possível determinação. Além disso, diferentes avaliações experimentais foram efetuadas para validar e avaliar as diferentes proposições.

A Figura 1.1 apresenta uma visão geral em relação aos desafios de pesquisa explorados por esta tese, de forma a destacar seus principais objetivos e contribuições:

- Inicialmente, para a construção de uma rede social acadêmica é fundamental analisar as interações entre pesquisadores. Nesta tese, focaliza-se na modelagem de relações de coautoria entre pesquisadores para tentar identificar os relacionamentos em pesquisa entre os mesmos. Para tanto, faz-se uso de informações sobre publicações dos autores que podem ser obtidas de diferentes fontes, por exemplo, bibliotecas digitais. Modelar essa rede e "ponderar" adequadamente essas relações para representar a importância dessas ligações constituem um grande desafio. A consideração em relação à quantidade de publicações aliada a aspectos temporais para a determinação de métricas, nesse sentido, pode ser fundamental para diferenciar relações mais "estabelecidas" (em frequência e atualidade).
- Uma análise mais aprofundada dessas relações pode ser utilizada para definir indicadores de qualidade associados à avaliação de grupos de pesquisadores, com base nas interações entre eles e determinar métricas para sua estimativa. Mais ainda, mostrando-se que tais métricas podem estimar/avaliar adequadamente a qualidade dos grupos, mostra-se que as interações e colaborações entre pesquisadores podem ser muito importantes no contexto acadêmico.
- Nesse cenário, abordagens considerando o contexto social, especificamente na área de recomendação de colaboradores, podem ser de grande valia para orientar e incentivar interações entre pesquisadores, visando também um incremento de qualidade para o grupo através de novas e intensificadas colaborações. Dessa forma, esta tese também explora indicadores importantes de serem considerados e propõe uma função de recomendação que agrega diferentes indicadores e pode gerar a recomendação detalhada de colaborações, incluindo a intensificação de colaborações anteriormente estabelecidas.

1.2 Organização do texto

O restante da tese está organizado como segue.

- No Capítulo 2, é apresentada uma revisão bibliográfica que aborda os assuntos permeados por esta tese. Esse capítulo inclui a apresentação de alguns conceitos envolvidos nas áreas de estudo e análise de redes sociais, avaliação de qualidade no contexto acadêmico e sistemas de recomendação. Também são apresentados, discutidos e comparados alguns trabalhos relacionados a esta tese.

- No Capítulo 3, é apresentada em detalhes a abordagem proposta nesta tese em relação à análise e avaliação de grupos de pesquisadores no contexto acadêmico. Esse capítulo inclui a apresentação do coeficiente de Gini aplicado à análise de redes sociais e as possíveis métricas associadas a indicadores de qualidade, considerando a questão social para ranquear grupos de pesquisadores. Nesse caso específico, as métricas são aplicadas a programas de Pós-graduação, mas poderiam ser aplicadas a quaisquer grupos de pesquisadores. Além disso, é apresentada uma avaliação experimental para validar e avaliar a abordagem proposta.

- No Capítulo 4, é apresentada a abordagem proposta para a recomendação de colaborações no contexto acadêmico. Esse capítulo apresenta alguns conceitos envolvidos e uma visão geral da abordagem proposta. Além disso, os indicadores definidos e as métricas correspondentes, bem como a função de recomendação, são apresentados em detalhes. Alguns refinamentos pela consideração de aspectos temporais, para ponderação dos vínculos relacionais entre pesquisadores, também são propostos e discutidos. Esse capítulo ainda inclui a apresentação de uma ampla avaliação experimental efetuada.

- Por fim, no Capítulo 5, são apresentadas as considerações finais, destacando as contribuições e os resultados obtidos por esta tese. Esse capítulo inclui a apresentação das publicações resultantes dos trabalhos realizados, no decorrer do doutorado, e discute alguns trabalhos futuros.

2 FUNDAMENTAÇÃO CONCEITUAL

Este capítulo apresenta um estudo bibliográfico relativo aos assuntos permeados por esta tese. Nas próximas seções, são apresentados: conceitos envolvidos nas áreas de estudo e análise de redes sociais (seção 2.1); avaliação de qualidade no contexto acadêmico (seção 2.2) e sistemas de recomendação (seção 2.3). Além disso, nessas seções, são apresentados e discutidos os principais trabalhos relacionados e os mesmos são comparados em relação à abordagem desta tese (seção 2.4).

2.1 Redes Sociais

A análise de Redes Sociais (*Social Network Analysis* - SNA) tem por base a assunção da importância dos relacionamentos entre unidades de interação. A perspectiva de redes sociais compreende teorias, modelos e aplicações que são expressos em termos de processos e conceitos relacionais. Dessa forma, as relações definidas pelas ligações entre unidades é um componente fundamental da teoria de redes.

O crescente interesse e uso de análises de redes sociais contribuiu para que fossem estabelecidos os princípios centrais que norteiam esse tipo de análise. Esses princípios distinguem análises de redes sociais de outras abordagens de pesquisa. Em adição ao uso de conceitos relacionais, Wasserman e Faust (1994) enfatizam o seguinte sobre redes sociais:

- Atores e suas ações são considerados como interdependentes, ao invés de independentes;
- Vínculos relacionais entre atores são vistos como canais para transferência de recursos;
- Modelos de rede focalizam os indivíduos em um ambiente estrutural da rede, para prover oportunidades ou restrições para ação individual;
- Modelos de rede conceituam a estrutura como estabelecedora de padrões de relações entre os atores.

O interesse acadêmico e do público em geral em redes sociais tem crescido rapidamente. A Figura 2.1 mostra que publicações em Ciência da Computação tendo "redes sociais" como um conceito chave aceleraram, na última década, de forma quase exponencial. Esse gráfico foi construído considerando publicações indexadas na DBLP até Julho de 2011. O gráfico apresenta o número total de publicações que contêm ambas as sequências de caracteres "Social" e "Network" (como termos ou *substrings*) em seus títulos, por

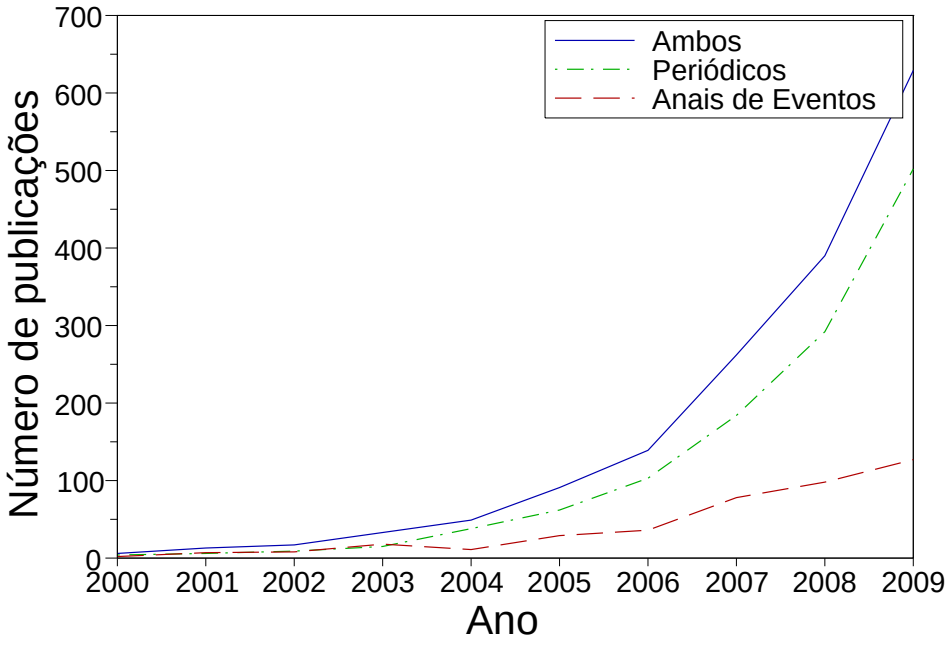


Figura 2.1: Crescimento de publicações sobre Redes Sociais em Ciência da Computação.

ano. Tal gráfico apresenta ainda os resultados separados nas seguintes categorias que representam os veículos de publicação considerados: (i) anais de eventos; (ii) periódicos e (iii) ambos. Foi considerado o período de 2000 a 2009, para não haver distorções nos resultados apresentados, devido ao atraso na indexação de publicações existente na DBLP, pois algumas podem levar mais de um ano para serem indexadas.

O crescimento de pesquisas e estudos sobre Redes Sociais (*Social Networks* - SN), nos tempos atuais, também se deve ao fato de o conceito desse tipo de estrutura ter passado a fazer parte de uma das aplicações mais utilizadas atualmente na Internet: as redes de relacionamento. Alguns exemplos de redes de relacionamento são: Facebook, QZone[1], V Kontakte[2], dentre outros. Esse tipo de rede passou a fazer parte da vida de muitas pessoas. A edição do "Mapa das Redes Sociais no Mundo" (*World Map of Social Networks*), apresentada na Figura 2.2, mostra as redes de relacionamento mais populares, por países, em Dezembro de 2011, segundo a *Alexa & Google Trends for Websites*. Em Fevereiro de 2012, a rede líder (Facebook) já possuía mais de 845 milhões de usuários ativos[3]. É importante analisar também que há um crescimento na difusão desse tipo de rede de relacionamento, sendo que esse já está sendo utilizado em todos os continentes. Um outro exemplo do uso do conceito de redes sociais pode ser a rede de coautoria de trabalhos entre pesquisadores. Nesse tipo de rede, também se tem o conceito de rede de relacionamentos, onde os indivíduos da rede são pesquisadores e os relacionamentos podem ser os trabalhos desenvolvidos em conjunto. Esse último pode ser considerado um exemplo de uma rede acadêmica representando as relações de colaboração, através de coautorias, entre pesquisadores.

Um consenso empírico sobre redes sociais é em relação ao fenômeno conhecido como "mundo pequeno". Tal fenômeno refere-se ao fato de que o mundo parece "pequeno"

[1]QZone: http://qzone.qq.com/

[2]V Kontakte: http://vkontakte.ru/

[3]Facebook has over 845 million users: http://www.zdnet.com/blog/facebook/facebook-has-over-845-million-users/8332

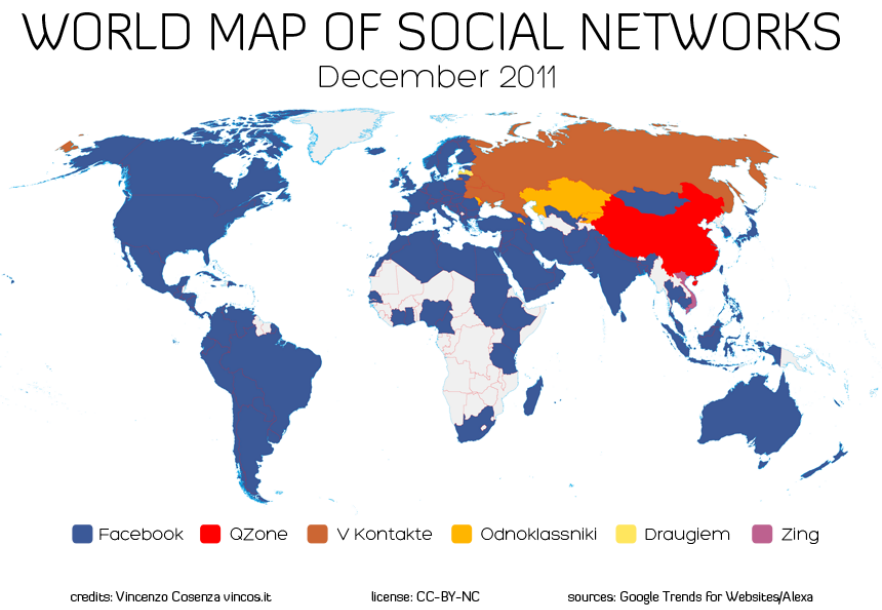


Figura 2.2: Mapa das Redes Sociais no Mundo em Dezembro de 2011 (fonte: http://vincos.it/world-map-of-social-networks/).

quando se pensa em quão pequenos são os caminhos que conectam quaisquer dois indivíduos em uma rede de relacionamentos. Nesse sentido, surgiu também a hipótese dos "seis graus de separação". Um dos primeiros experimentos realizados para comprovar esse fenômeno foi efetuado por Milgram (1967). Nesse experimento, Stanley Milgram escolheu ao acaso, inicialmente, 296 voluntários, na tentativa de encaminhar, cada um, uma carta a uma pessoa alvo designada (no subúrbio de Boston). As cartas deveriam chegar ao usuário alvo através das relações de amizade. Cada indivíduo, mesmo sem conhecer o usuário alvo diretamente, foi estimulado a encaminhar uma carta a outra pessoa que possivelmente poderia ajudar a descobrir alguém que conhecesse o usuário alvo. Dessa forma, cerca de um terço das cartas chegou ao seu destino, sendo que, em média, foram necessários seis encaminhamentos para uma carta chegar ao seu destino final. Os resultados desse experimento mostraram evidências da existência de caminhos curtos na rede de amizade global.

Especificamente em redes de colaboração acadêmicas, o fenômeno do "mundo pequeno" foi também estudado e descobriram-se caminhos muito curtos, nesse tipo de rede, dentro de comunidades de pesquisa. Na área da matemática, por exemplo, um experimento foi conduzido considerando o matemático Paul Erdös, que publicou cerca de 1.500 artigos em sua carreira, como uma figura central na estrutura colaborativa desta área. Um experimento foi efetuado para determinação do número de *Erdös* (distância entre pesquisadores e Paul Erdös). A maioria dos matemáticos, e até alguns pesquisadores de outras áreas, tinham um número de *Erdös* extremamente pequeno, no máximo 4 ou 5. Segundo Easley e Kleinberg (2010), isso demonstra que o mundo da ciência é verdadeiramente pequeno nesse sentido. Em uma pesquisa divulgada em novembro de 2011, cientistas do Facebook e da Universidade de Milão, considerando os dados da rede de relacionamento Facebook, estimaram que o grau de separação médio entre quaisquer dois usuários, em senso global, é de aproximadamente 4,74[4].

[4]Separating You and Me? 4.74 Degrees: http://www.nytimes.com/2011/11/22/technology/between-

A importância crítica do desenvolvimento de métodos para análises de redes sociais é o fato de que a unidade de análise em uma rede social não é um indivíduo, mas uma entidade consistindo de coleções de indivíduos e ligações entre eles. A análise de redes sociais é inerentemente um esforço interdisciplinar. Os conceitos de análises de redes sociais são desenvolvidos a partir de um propício encontro entre teoria e aplicação social com matemática formal, estatística e metodologia computacional (WASSERMAN; FAUST, 1994). Os pioneiros em análises de redes sociais vêm da sociologia, psicologia social e antropologia, sendo que é atribuído a Barnes (1954) o primeiro uso do termo "rede social".

Os métodos de análises de redes sociais provêm declarações formais sobre processos e propriedades sociais. Além disso, esses conceitos devem ser definidos de modo preciso e consistente. Uma vez definidos, os mesmos podem fornecer novas perspectivas sobre o mundo social (FREEMAN, 1984).

A análise das redes sociais parte de duas grandes visões do objeto de estudo: as redes inteiras (*whole networks*) e as redes personalizadas (*ego-centered networks*). Na primeira visão, o foco é a relação estrutural da rede com o grupo social. De acordo com essa visão, as redes são assinaturas de identidade social, sendo que o padrão de relações entre os indivíduos mapeia as preferências e características dos próprios envolvidos na rede (WATTS, 2003). Na segunda visão, o foco está no papel social de um indivíduo, que pode ser compreendido não apenas através dos grupos (redes) a que ele pertence, mas igualmente, através das posições que ele ocupa dentro dessas redes. A diferença entre os dois focos está no *corpus* da análise escolhida pelo pesquisador (RECUERO, 2005).

2.1.1 Conceitos em Análises de Redes

A seguir, são comentados alguns conceitos utilizados na análise de redes, que são fundamentais na discussão sobre redes sociais, incluindo: atores, vínculos relacionais, díade, tríade, subgrupo, grupo, relação e rede social. Essas definições são baseadas em (WASSERMAN; FAUST, 1994).

Atores. Atores são entidades que possuem ligações sociais modeladas pela rede social. Atores podem ser indivíduos, corporações ou unidades sociais coletivas, por exemplo: pessoas em um grupo, departamento ou dentro de uma corporação, etc. O uso do termo "ator" não implica que essas entidades necessariamente tenham a habilidade de "agir". Muitas aplicações de redes sociais focalizam coleções de atores que são todos do mesmo tipo, por exemplo, pessoas em um grupo de trabalho. Entretanto, existem redes formadas por atores que são de tipos e níveis conceituais distintos ou de diferentes conjuntos.

Vínculos relacionais. Os atores são ligados a outros através de vínculos sociais. O intervalo e o tipo de vínculos possíveis são bastante extensos. A definição característica de um vínculo é estabelecer uma ligação entre um par de atores. Alguns exemplos de vínculos sociais incluem avaliação de uma pessoa pela outra (expressa pelo nível de amizade, ligação ou respeito), transferência material de recursos (transações de negócios), associação ou afiliação (participantes de um mesmo evento social ou do mesmo clube social), interações comportamentais (conversar juntos, troca de mensagens), movimentação entre lugares ou de *status* (migração social ou mobilidade física), conexão física (estrada, rio ou ponte conectando dois pontos), relações formais (por ex., autoridade), relações biológicas (ancestral ou descendente), etc.

---

you-and-me-4-74-degrees.html?_r=4&scp=1&sq=4.74&st=cse

Díade. Uma díade consiste de um par de atores e a possível ligação entre eles. No nível básico, uma ligação ou relação estabelece um vínculo entre dois atores. O vínculo é inerentemente uma propriedade do par e, portanto, não é considerado como pertencente simplesmente a um ator individual. Muitos tipos de análises de redes são concebidos com o entendimento dos vínculos entre pares. Essas abordagens adotam a díade como uma unidade de análise. Análises de díades centram-se na propriedade de relações entre pares, tais como: se vínculos são recíprocos ou não; se tipos específicos de múltiplos relacionamentos tendem a ocorrer juntos. A díade é frequentemente a unidade básica para análises estatísticas sobre redes sociais.

Tríade. Uma tríade é um subconjunto de três atores e um possível vínculo entre eles (três potenciais pares de vínculos). Muitos métodos e modelos importantes em redes sociais baseiam-se em tríades. Alguns tipos de análises que poderiam ser estudados: se uma tríade é transitiva (se o ator i gosta do ator j, e o ator j, por sua vez, gosta do ator k, então o ator i também gostará do ator k) ou se uma tríade é balanceada (se os atores i e j gostam um do outro, então i e j são similares em suas avaliações de um terceiro ator k; e se i e j não se gostam, então eles diferem nas suas avaliações de um terceiro ator k).

Subgrupo. O conceito de subgrupo está associado a um subgrupo de atores com qualquer subconjunto de atores e todos os vínculos entre eles. Localizar e estudar subgrupos, utilizando critérios específicos, tem sido uma importante preocupação em análises de redes sociais.

Grupo. Um grupo é uma coleção de todos os atores sobre os quais vínculos podem ser estabelecidos. Um grupo consiste de um conjunto finito de atores que, por razões conceituais, teóricas ou empíricas, é tratado como um conjunto finito de indivíduos sobre os quais métricas de rede são feitas.

Relação. Relação é a coleção de vínculos de um tipo específico entre membros de um grupo. Por exemplo: um conjunto de amizades entre pares de alunos em uma turma ou vínculos diplomáticos mantidos entre pares de nações no mundo são vínculos que definem relações. Uma relação se refere à coleção de vínculos de um dado tipo estabelecida entre pares de atores de um específico conjunto de atores.

Rede Social. Tendo em vista os conceitos definidos anteriormente, uma Rede Social consiste de um conjunto ou conjuntos finitos de atores e a relação ou relações definidas entre eles. A presença de informação relacional é uma característica de definição crítica em uma rede social.

A representação de uma rede social fictícia é apresentada na Figura 2.3. Nessa, alguns exemplos dos conceitos descritos acima são explicitamente apontados (ator, vínculo relacional, díade, tríade e subgrupo). Considere que essa é uma rede acadêmica, onde cada círculo representa um pesquisador e a existência de pelo menos uma coautoria entre pares de pesquisadores é representada por uma linha ligando-os. Dessa forma, além dos conceitos explicitamente indicados, o conjunto de todos os círculos representa o grupo, conjunto de pesquisadores, e o conjunto de todas as linhas representa a relação, coautorias entre pesquisadores.

2.1.2 Métricas em análises de redes

Esta seção apresenta algumas das métricas tradicionais aplicadas em Análise de Redes Sociais (FREEMAN, 1979; HOSER et al., 2006; MARSDEN, 2002; NEWMAN, 2001, 2003) seguidas de um exemplo completo que apresenta valores das mesmas para uma

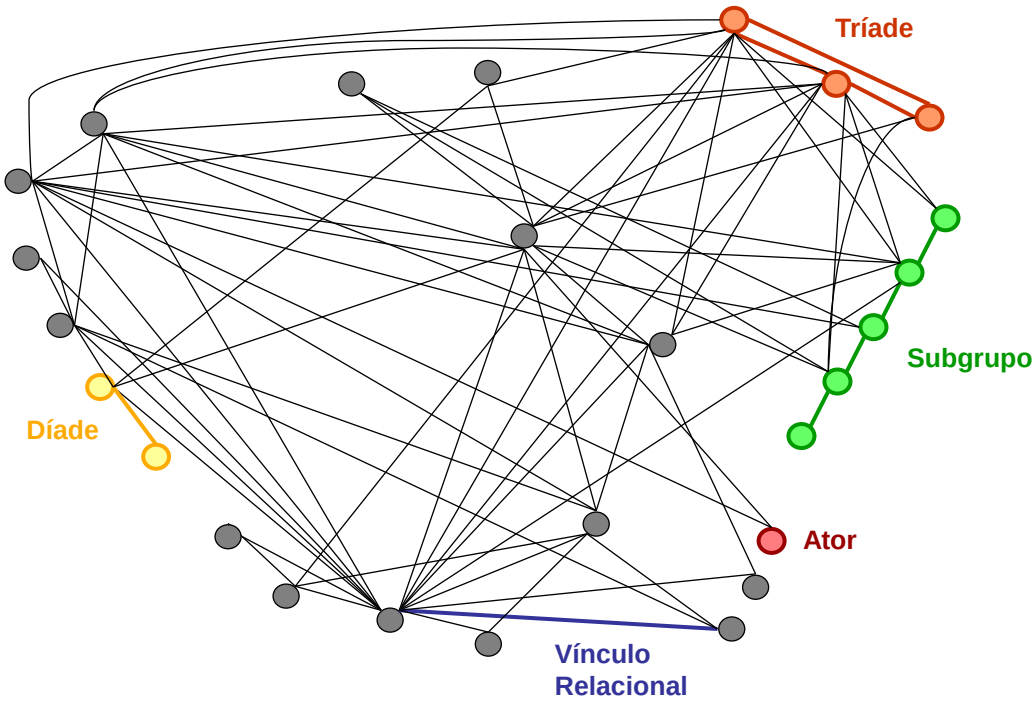


Figura 2.3: Representação de uma rede social fictícia para a demonstração de exemplos dos conceitos básicos de SNA.

rede social fictícia. O modo tradicional de representar uma rede social é através de um grafo $G := (\aleph, \xi)$, com nós (vértices) $n \in \aleph$ e arestas (ligações, conexões) $e \in \xi$. Além disso, os nós representam os atores da rede e as arestas seus vínculos relacionais. Para a apresentação das equações, o número total de nós, na rede social sendo analisada, é representado como N; e $e(n_i, n_k)$ retorna 1 quando existe uma aresta entre os nós indicados (n_i e n_k) e 0 (zero) caso contrário.

Grau de centralidade. O conceito de grau de centralidade (*degree centrality*) presume que um nó que tem muitas conexões é considerado importante, enquanto que um nó sem conexões é considerado irrelevante. Esse grau reflete a atividade relacional direta de um nó (FREEMAN, 1979). O grau de centralidade de um nó é calculado como o número de vínculos diretos, arestas imediatas, que o envolvem. A Equação 2.1 apresenta o cálculo do grau de centralidade de um nó n_i, chamado $dc(n_i)$. Se a rede é um grafo não-direcionado, ou seja, a conexão entre dois nós é não-direcionada, a métrica é chamada apenas de grau. Se a rede é um grafo direcionado, a métrica é categorizada em grau de entrada (*in-degree*) e grau de saída (*out-degree*) de acordo com a direção dos relacionamentos sendo analisados.

$$dc(n_i) = \sum_{k=1}^{N} e(n_i, n_k) \tag{2.1}$$

Densidade. A densidade (*density*) de uma rede é definida com base no grau de centralidade, sendo calculada como o número de arestas existentes dividido pelo número de todas as possíveis arestas dessa rede. Uma rede inteiramente conectada tem densidade igual a 1. O número de todas as possíveis arestas muda de acordo com o tipo de grafo que descreve a rede. Esse conceito não é útil quando múltiplas arestas são permitidas ou quando as arestas possuem pesos atribuídos, porque o número total de possíveis conexões

não pode ser avaliado. Se a rede é um grafo não-direcionado onde somente uma aresta é permitida, o número possível de conexões entre cada dois nós é 1, e a Equação 2.2 pode ser usada para calcular a densidade (nomeada como d) do grafo G representando a rede. Nessa equação, o número total de arestas é calculado pela soma do grau de centralidade ($dc(n_i)$) de todos os nós dividida por 2 (para que as arestas entre dois nós sejam contadas apenas uma vez), e o número de todas as possíveis arestas é calculado para uma rede não-direcionada como $(N(N-1)/2)$ onde N representa o número total de nós.

$$d(G) = \frac{\sum_{i=1}^{N} dc(n_i)}{N(N-1)} \tag{2.2}$$

Coeficiente de clusterização. O coeficiente de clusterização (*clustering coefficient*) de um nó n_i, nomeado $cc(n_i)$, é calculado como o número de arestas existentes entre vizinhos de n_i dividido pelo número total de todas as possíveis arestas entre os vizinhos do nó n_i. O coeficiente de clusterização de um nó objetiva determinar a densidade de arestas estabelecidas entre os vizinhos de um nó. O conceito de "transitividade" é aplicado descrevendo simetria de interação entre triplas de nós (análise das tríades da SN). Três nós n_1, n_2 e n_3 são transitivos se para n_1 é conectado a n_2, o qual é conectado a n_3 implica na conexão de n_1 e n_3. A transitividade entre triplas de nós é calculada como o número de triplas que são transitivas dividido pelo número de triplas que têm potencial de serem transitivas (caminhos de tamanho 2). O coeficiente de clusterização global (*overall clustering coefficient*) de uma rede (Equação 2.3), nomeado occ, é calculado como a média do coeficiente de clusterização de todos os seus nós. Para o coeficiente de clusterização de uma rede, também pode ser usado um coeficiente de clusterização ponderado (*weighted clustering coefficient*) (Equação 2.4), nomeado wcc, que é calculado pela média ponderada do coeficiente de clusterização de todos os nós, cada um ponderado pelo seu grau de centralidade (dc).

$$occ(G) = \frac{1}{N} \sum_{i=1}^{N} cc(n_i) \tag{2.3}$$

$$wcc(G) = \frac{\sum_{i=1}^{N} dc(n_i) cc(n_i)}{\sum_{i=1}^{N} dc(n_i)} \tag{2.4}$$

Coeficiente gigante. O coeficiente gigante (*giant coefficient*) é calculado com base no tamanho do componente gigante de uma rede. O componente gigante (*giant component*), também chamado como componente principal (*main component*), é o componente conectado com o maior número de nós. O componente principal (*main component* - MC) é um subgrafo do grafo G representando a rede (ou pode ser o próprio grafo G no caso de uma rede totalmente conectada). O coeficiente gigante (Equação 2.5), nomeado como gc, é calculado como o tamanho (número de nós) do componente principal dividido pelo número total de nós da rede sendo analisada. Esse valor representa o percentual de nós que são parte do componente gigante.

$$gc(G) = \frac{1}{N} \sum_{i=1}^{N} mc(n_i) \tag{2.5}$$

onde:

$$mc(n_i) = \begin{cases} 1, & \text{se } (n_i \in MC) \\ 0, & \text{caso contrário} \end{cases} \tag{2.6}$$

Grau de Proximidade. O grau de proximidade (*closeness centrality*) descreve o nível de liberdade dos nós em uma rede ou simplesmente sua capacidade para ação independente dentro da rede. O grau de proximidade de um nó é definido pelo inverso da soma de suas distâncias (menor caminho) para todos os outros nós.

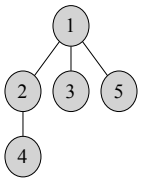
Grau de Intermediação. O grau de intermediação (*betweenness centrality*) descreve a localização global de um nó na rede. Esse grau reflete a localização intermediária de um nó ao longo de relacionamentos indiretos ligando outros nós. O grau de intermediação de um nó é calculado pelo número de menores caminhos entre quaisquer dois nós que passam através de um determinado nó (pode ou não ser normalizado). Essa métrica provê frequentemente um alto grau de informação, como descrever a localização de um nó no grafo em senso global, enquanto o grau de centralidade considera apenas os vizinhos diretos.

Diâmetro. O diâmetro (*diameter*) é outra métrica associada à distância no grafo. Essa métrica é calculada como o valor máximo obtido entre todos os menores caminhos avaliados entre pares de nós do grafo, ou seja, maior distância entre quaisquer pares de nós pertencentes ao grafo. Geralmente redes sociais têm um diâmetro pequeno. O diâmetro bem como a densidade são métricas geralmente utilizadas para comparação entre redes.

Exemplo de cálculo das métricas. A Figura 2.4 mostra um exemplo para todas as métricas de SNA apresentadas nesta seção. Especificamente, a Figura 2.4(a) ilustra uma rede social (não direcionada) simples composta por apenas cinco atores, e a Figura 2.4(b) mostra os valores das métricas estimadas para essa rede. Nesse exemplo, a rede tem uma única componente conexa. Normalmente, quando não existe uma única componente, a componente principal, componente conexa com o maior número de nós, deve ser usada para estimativa das métricas que consideram a rede globalmente, não calculadas em relação a um único nó. As primeiras três métricas são calculadas para nós específicos. Aqui, foi escolhido o nó numerado como 1 (#1). O grau de centralidade para o nó #1 é 3, referenciando o número de seus vínculos relacionais (arestas). O valor do grau de proximidade é 0,2, e foi calculado por 1 dividido pela soma dos menores caminhos entre o nó #1 e todos os outros nós, ou seja, (1/(3 menores caminhos de tamanho 1 + 1 menor caminho de tamanho 2)). O grau de intermediação, não-normalizado, é 5, porque existem 5 menores caminhos entre pares de nós que passam através do nó #1 (ou seja, #2 e #3, #2 e #5, #3 e #4, #3 e #5, e #4 e #5). As quatro últimas métricas são estimadas para a SN como um todo. O valor de densidade é 0,4, ou seja, o número total de arestas existentes (4) dividido pelo número total de possíveis conexões (10). O valor de diâmetro é 3, ou seja, o valor do máximo menor caminho (nesse caso, ocorre entre os nós #3 e #4 ou #4 e #5). O coeficiente gigante desta rede é 1, já que esta possui uma única componente conexa, que corresponderá à componente principal ((5 nós pertencentes a maior componente conexa)/(5 nós correspondentes a rede sendo analisada)). Já o coeficiente de clusterização desta rede é 0, tanto o ponderado quanto o global, uma vez que o coeficiente de clusterização de todos os nós é 0 porque nenhum dos vizinhos diretos de um nó possuem arestas entre si.

2.2 Avaliação de qualidade no contexto acadêmico

Esta seção apresenta questões relacionadas à avaliação de qualidade/impacto no contexto acadêmico. Inicialmente, são apresentadas métricas tradicionais da área de bibliometria (seção 2.2.1), a seguir, é abordada a qualidade visando um contexto social (seção



(a)

Métricas	Valores
Grau de centralidade (#1)	3
Grau de proximidade (#1)	0,2
Grau de intermediação (#1)	5
Densidade	0,4
Diâmetro	3
Coeficiente gigante	1
Coeficiente de clusterização	0

(b)

Figura 2.4: Exemplos de (a) uma rede social simples e (b) respectivos valores das métricas.

2.2.2) e, por fim, são apresentadas outras análises com o uso do Coeficiente de Gini (seção 2.2.3).

2.2.1 Métricas em bibliometria

A determinação de métricas para avaliação de qualidade no contexto acadêmico sempre esteve muito relacionada a dados de bibliometria, especificamente a estatísticas de citações. Nesse sentido, esta seção apresenta brevemente dois trabalhos: fator de impacto e índice h (apresentados nas seções 2.2.1.1 e 2.2.1.2). Eles podem ser considerados marcos nesta área, já que muitos outros trabalhos foram desenvolvidos utilizando-os como base. Os trabalhos originados abordam a especificação ou subsídios para a especificação de métricas para determinar a qualidade/nível ou o impacto de, por exemplo, veículos de publicação (conferências e periódicos), artigos científicos e pesquisadores.

2.2.1.1 Fator de Impacto

Segundo Gowrishankar et al. (1999), fatores de impacto de periódicos foram introduzidos nos anos 70 para "ranquear" diferentes periódicos, através de análises de citações e surgiram da necessidade de instituições científicas estarem sendo julgadas cada vez mais pela qualidade dos periódicos nos quais sua equipe publica artigos. Essa "qualidade" de um periódico é habitualmente representada pelo seu fator de impacto, que visa mensurar uma aproximação da importância de um periódico em sua área.

O Fator de Impacto (*Impact Factor* - IF) foi desenvolvido por Eugene Garfield, o fundador do *Institute for Scientific Information* (ISI), hoje parte da Thomson Corporation. Segundo Garfield (2006), a primeira vez que ele mencionou a ideia de "fator de impacto" foi em (GARFIELD, 1955), onde destacou a necessidade de avaliar a importância relativa de periódicos científicos, com base em informações sobre os índices de citações.

O IF é calculado a cada ano pela Thomson Scientific para aqueles periódicos que ela indexa, e o fator e índices são publicados no *Journal Citation Reports* (JCR). O IF de um periódico é calculado com base em um período de três anos, visando ser uma aproximação do número médio de citações em um ano. São computadas as citações daqueles artigos do periódico que foram publicados durante os dois anos anteriores.

O cálculo do IF de um periódico em determinado ano é realizado através da Equação 2.7.

$$IF = A/D \tag{2.7}$$

onde:

- IF = fator de impacto em Y;
- Y = ano de análise;
- A = citações em Y para artigos publicados entre Y-2 e Y-1;
- D = número de artigos publicados entre Y-2 e Y-1.

Deve-se observar que o fator de impacto em Y é publicado em Y+1, isso porque tal índice não pode ser calculado até que todas as publicações no ano Y tenham sido recebidas.

Algumas informações importantes a serem consideradas são apresentadas a seguir[5]: (i) a ISI exclui certos tipos de artigos do denominador, tais como: novos itens, correspondência e errata; (ii) novos periódicos, que são indexados por seu primeiro número de publicação, recebem um fator de impacto depois de dois anos completos de indexação; nesse caso, as citações do ano anterior ao Volume 1, e o número de artigos publicados no ano anterior ao Volume 1 são zero; (iii) periódicos que são indexados iniciando com um outro volume que não o primeiro volume não têm um fator de impacto publicado até completos três anos de dados conhecidos; (iv) publicações anuais e outras irregulares que não publicam itens em um ano particular afetam a contagem; (v) o *Journal Citation Reports* (JCR) inclui uma tabela do *ranking* relativo de periódicos pelo fator de impacto por área específica.

Uma das variações proposta sobre o fator de impacto, nomeada *Impact Fator Revised* (IF_r), leva em consideração a exclusão de auto-citações[6]. Essa é uma característica importante de ser analisada já que auto-citações não atestam o impacto real de uma publicação frente à comunidade científica.

2.2.1.2 Índice h

O índice h (h-index) (HIRSCH, 2005), também conhecido como índice de *Hirsch* ou número de *Hirsch*, é um índice proposto originalmente para quantificar o impacto e a relevância do resultado da pesquisa científica de um pesquisador. Algumas informações úteis para esse tipo de cálculo, segundo Hirsch (2005), são o número de artigos publicados em n anos, o número de citações de cada artigo, os periódicos onde os artigos foram publicados, seus parâmetros de impacto, etc. Porém, essas informações podem ser avaliadas com diferentes critérios por diferentes pessoas. Hirsch propôs um único número, o índice h, como um modo particularmente simples e útil de caracterizar o resultado da pesquisa científica de um pesquisador. O índice h é calculado com base no número de artigos publicados em n anos e no número de citações de cada artigo publicado. Esse número indica que um cientista tem índice h, se h de seus Np artigos têm pelo menos h citações cada e os outros $(Np - h)$ artigos têm $\leq h$ citações cada.

Um autor, com um dado índice h, tem publicado h artigos cada qual tendo sido citado por outros pelo menos h vezes. Então, o índice h reflete ambos: número de publicações e número de citações por publicação, como pode ser visto na Figura 2.5. A Figura 2.5 apresenta um gráfico esquemático do número de citações *versus* o número de artigos, sendo que os artigos foram plotados ordenados decrescentemente pelo número de citações. A curva apresentada é uma aproximação do comportamento esperado nesse gráfico. A intersecção da linha de 45^o com a curva corresponde a h. Nesse ponto de intersecção, têm-se o

[5]Impact Factor - Wikipedia: http://en.wikipedia.org/wiki/Impact_factor

[6]Impact Factor: http://scientific.thomsonreuters.com/free/essays/journalcitationreports/impactfactor/

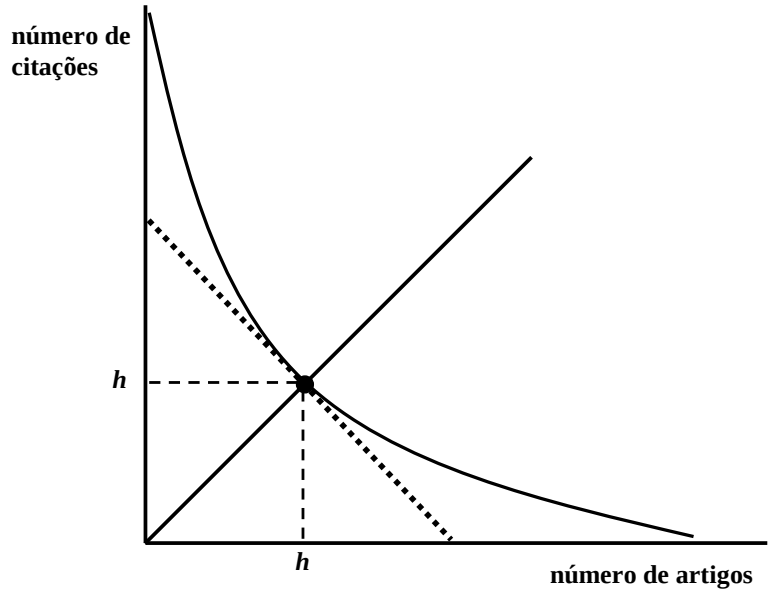


Figura 2.5: Gráfico esquemático do número de citações *versus* número de artigos, com artigos ordenados decrescentemente por citação. A intersecção da linha de 45^o com a curva corresponde a h (adaptado de (HIRSCH, 2005)).

número de citações e o número de artigos, ambos iguais e com valor h. Maiores detalhes sobre o formalismo matemático envolvido na definição do índice h podem ser obtidos em (HIRSCH, 2005).

O índice h é designado para prover uma medida simples do número total de citações por publicações. O índice trabalha adequadamente somente para comparação de cientistas da mesma área; convenções de citações diferem amplamente entre as diversas áreas do conhecimento. O índice h serve como uma alternativa às mais tradicionais métricas de fator de impacto de periódicos na avaliação do impacto de um trabalho de um pesquisador particular[7].

Embora no cálculo do índice h tradicional as auto-citações não sejam descartadas, o efeito desse problema é muito menor para o cálculo desse índice, já que auto-citações de artigos com $< h$ citações serão irrelevantes, assim como auto-citações para artigos com muito mais do que h citações (HIRSCH, 2005). Além disso, o autor apresenta uma maneira de calcular o índice h levando em consideração o problema das auto-citações.

Índices como o índice h podem ser úteis para ranquear pesquisadores de determinadas áreas. Um exemplo é o *ranking* de pesquisadores da área da Ciência da Computação que vem sendo disponibilizado por Jens Palsberg (UCLA)[8], o qual é calculado com base em dados obtidos do Google Scholar. Tal *ranking* apresenta uma lista parcial dos pesquisadores, ordenados decrescentemente pelos seus respectivos valores de h, que possuem índice $h \geq 40$. O valor máximo de índice h encontrado para um pesquisador da área da Ciência da Computação, em 06 de janeiro de 2012, foi de 110. Esse valor máximo pode variar, consideravelmente, dependendo da área de pesquisa que está sendo considerada.

Existem algumas ferramentas automatizadas para o cálculo do índice h, como por

[7]Hirsch number - Wikipedia: http://en.wikipedia.org/wiki/Hirsch_number
[8]The h Index for Computer Science: http://www.cs.ucla.edu/˜palsberg/h-number.html

exemplo: *QuadSearch*[9], *Publish or Perish*[10] e *scHolar index*[11]. As ferramentas citadas calculam o valor de h estimado com base em dados obtidos do Google Scholar. *Publish or Perish* calcula diversas estatísticas e apresenta não só o cálculo do índice h tradicional, calculado para pesquisadores, mas também calcula o índice h estimado para conferências e periódicos, seguindo o mesmo princípio do cálculo do índice h para pesquisadores.

Algumas vantagens do índice h são descritas na literatura (EGGHE, 2006), como: (i) ser um número único e simples incorporando ambos escores de publicações (quantidade) e de citações (qualidade ou visibilidade); (ii) ter uma vantagem sobre essas medidas únicas separadas e sobre medidas tais como "número de artigos significativos" (que é arbitrário) ou "número de citações de cada um dos artigos mais citados" (que de novo não é um simples número); (iii) ser robusto no sentido de ser insensível a um conjunto de artigos acidentalmente não citados, ou com baixo número de citações, e também para um ou vários artigos com número de citações excepcionalmente altas.

2.2.2 Qualidade e o contexto social

A determinação de indicadores de qualidade, no contexto acadêmico, esteve sempre muito centrada na área de bibliometria, especialmente, estatísticas de citações. Uma faceta ainda pouco explorada é o uso do contexto social como influenciador em um possível indicador de qualidade. Com o crescimento de estudos em redes sociais, essa situação começou a ser vislumbrada, mas ainda há pouco trabalho sendo desenvolvido nesse sentido. Dessa forma, uma das hipóteses exploradas nesta tese é justamente nesta direção: definir indicadores de qualidade considerando o contexto social e explorar as métricas de análise de redes sociais que possam auxiliar na estimação desses. Por exemplo, na métrica do grau de centralidade, quanto maior o número de conexões do indivíduo, maior a sua importância em relação à conectividade, que pode estar associada a uma maior "qualidade" do indivíduo estimada em relação a um indicador considerando o contexto social. Dessa forma, aliar métricas de análises de redes sociais e determinar novas, para exploração de indicadores, visando associar a "esfera social" à qualidade, pode ser muito importante nesse "mundo em crescente conexão".

Especificamente, no contexto acadêmico, a questão social pode ser explorada analisando as colaborações (coautorias) entre pesquisadores. Nesse caso, esse tipo de indicador de qualidade pode ser aplicado em diversas áreas do conhecimento, onde a tarefa de pesquisa científica é coletiva e não deve ser uma tarefa puramente individual. Assim, indicadores de qualidade relacionados à conectividade podem ser explorados tanto para "qualificar" o indivíduo quanto um grupo como um todo.

Trabalhos prévios em avaliação de qualidade dentro do contexto acadêmico objetivam avaliar pesquisadores, periódicos, conferências, instituições, dentre outros. Por exemplo, algumas abordagens foram desenvolvidas para avaliar outras questões considerando o grupo de pesquisadores envolvidos, tais como (SILVA et al., 2010; YAN; LEE, 2007). No entanto, muitas abordagens em avaliação de qualidade no contexto acadêmico consideram técnicas bibliométricas, especialmente estatísticas de citações. Uma discussão aprofundada sobre as vantagens e desvantagens do uso de estatísticas de citações é apresentada em (ADLER; EWING; TAYLOR, 2008). Exemplos de tais propostas baseadas em estatísticas de citações incluem aquelas para avaliar pesquisadores (EGGHE, 2006; HIRSCH, 2005; SILVA et al., 2010), instituições (REN; TAYLOR, 2007) e veículos de publica-

[9] QuadSearch: http://quadsearch.csd.auth.gr/index.php?lan=1&s=2
[10] Publish or Perish: http://www.harzing.com/pop.htm
[11] scHolar index: http://www-ihm.lri.fr/~roussel/moulinette/h/h.cgi

ção (HABIBZADEH; YADOLLAHIE, 2008; NICOLAISEN; FRANDSEN, 2008; YAN; LEE, 2007).

Um dos indicadores mais populares que considera técnicas bibliométricas é o índice h proposto por Hirsch (2005) (conforme apresentado na seção 2.2.1.2). Outras métricas para determinação de qualidade/impacto com base no índice h (chamados índices tipo h (EGGHE, 2010)) têm sido propostas. Alguns exemplos dessas, especificamente introduzidas para avaliar grupos de pesquisadores ou instituições, podem ser encontrados em (DA LUZ et al., 2008; MOLINARI; MOLINARI, 2008; RAAN, 2006).

Combinar diferentes critérios para obter a avaliação de qualidade/impacto é também utilizado (SOUTO; WARPECHOWSKI; OLIVEIRA, 2007). Note que a vantagem de utilizar um critério único é a simplicidade no processo de avaliação. Entretanto, não é fácil, e nem sempre é possível, descobrir uma única faceta do problema que gere resultados satisfatórios. Por outro lado, a utilização de múltiplos critérios focaliza a obtenção de resultados melhor "fundamentados" (embora, nem sempre, essa consideração de uma série de critérios tenda a levar os resultados a serem melhores).

Em geral, pesquisas são feitas através de colaborações envolvendo diferentes grupos de pesquisadores. Comunidades de pesquisa têm empregado análises de redes sociais para entender suas próprias interconexões, evolução e comportamentos (DING, 2011; MENEZES et al., 2009; WANG et al., 2010). Exemplos variam de comunidades consolidadas como a física (NEWMAN, 2001) e a matemática (BARABÁSI, 2002) até outras relativamente novas, tais como recuperação de informação (SMEATON et al., 2003). Portanto, analisar as colaborações entre pesquisadores, bem como quantificar esse comportamento de interações científicas, pode definir indicadores interessantes para a análise da qualidade global.

2.2.3 Outras análises com o uso do coeficiente de Gini

Esta tese estuda a possibilidade de análise de redes sociais, além de pelas métricas tradicionais presentes na literatura (ver seção 2.1.2), feita através do uso adaptado de métodos estatísticos para estudo de homogeneidade de distribuições, como o coeficiente de Gini. Por isso, tal coeficiente é apresentado a seguir.

O coeficiente de Gini é uma métrica de dispersão estatística proposta, em 1912, pelo estatístico italiano Corrado Gini no seu artigo "Variability and Mutability" (GINI, 1912). Esse coeficiente é comumente utilizado para avaliar desigualdade de distribuições de riqueza e renda, mas pode ser utilizado para outras distribuições (SUBRAMANIAN; KAWACHI, 2004).

O coeficiente de Gini é definido com base na curva de Lorenz. A curva de Lorenz foi criada por Max O. Lorenz em 1905. A curva de Lorenz é um gráfico que representa a distribuição cumulativa de uma função de densidade de probabilidade. Tal função é construída como um *ranking* dos membros da população dispostos em ordem crescente de riqueza, ou qualquer outra distribuição que se deseja estudar. Esses valores são denotados como $h_1 < h_2 < ... < h_{n-1} < h_n$ e considera-se a Equação 2.8, que calcula a fração de riqueza (distribuição cumulativa) correspondente à fração de pessoas $f_i = i/n, i = 1, ..., n$ na população.

$$\Phi\left(h_i\right) = \frac{1}{\left(\sum_{j=1}^{n} h_j\right)} \sum_{k=1}^{i} h_k \tag{2.8}$$

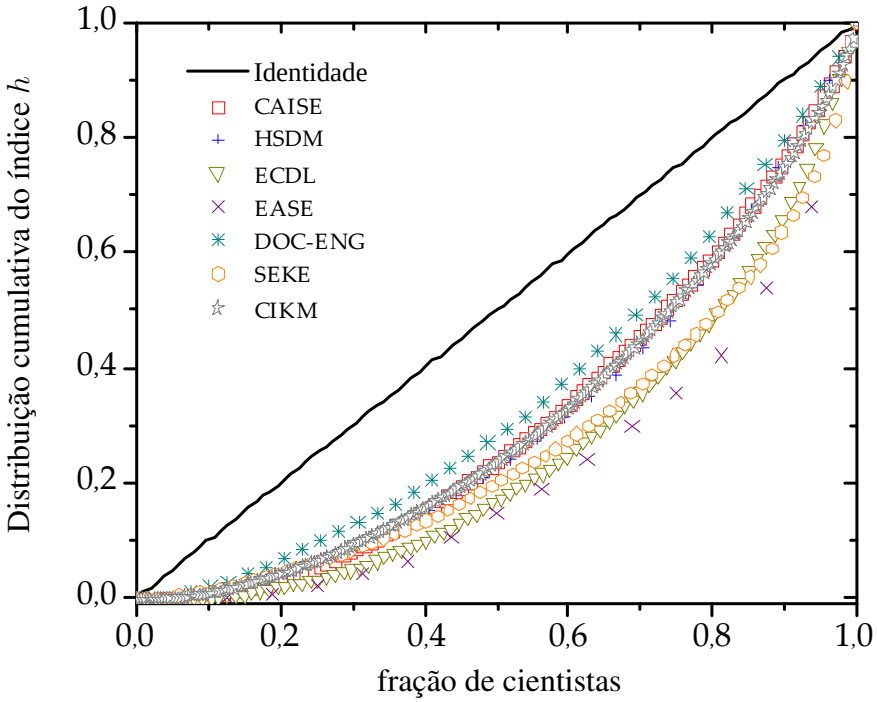


Figura 2.6: Curvas de Lorenz para as distribuições de índice h de pesquisadores em conferências de Engenharia de Software (adaptado de (SILVA et al., 2010)).

Em recentes experimentos, curvas de Lorenz têm sido empregadas para estudar a distribuição de índice h de pesquisadores em conferências (SILVA et al., 2010, 2012). A Figura 2.6 ilustra um típico gráfico para diferentes conferências analisadas nesses trabalhos.

Na Figura 2.6, o percentual de indivíduos é plotado no eixo x e o percentual dos valores da variável (na concepção original, a variável é a riqueza da população) no eixo y. A distribuição é perfeitamente igualitária quando todo indivíduo tem o mesmo valor de variável. Nesse caso, os $N\%$ do grupo de indivíduos terão sempre $N\%$ do valor de variável, e sua curva é $y = x$; chamada da linha de perfeita igualdade (identidade). Por outro lado, a distribuição perfeitamente desigual é aquela na qual somente um indivíduo tem todo o valor da variável. Nessa situação, a curva é $y = 0$ para todo $x < 100\%$, e $y = 100\%$ quando $x = 100\%$; chamada linha de perfeita desigualdade.

O coeficiente de Gini é calculado como o percentual da área entre a linha de perfeita igualdade e a curva de Lorenz observada em relação à área entre a linha de perfeita igualdade e a linha de perfeita desigualdade, como definido pela Equação 2.9.

$$g = 1 - 2\int_0^1 \Phi\left(h\right) dh \tag{2.9}$$

que é numericamente aproximado por uma fórmula trapezoidal, levando à Equação 2.10

$$g \cong 1 - \frac{\Phi\left(h_0\right) + \Phi\left(h_n\right)}{n} - \frac{2}{n}\sum_{k=1}^{n-1}\Phi\left(h_k\right) = 1 - \frac{1}{n}\sum_{k=1}^{n}\left[\Phi\left(h_k\right) + \Phi\left(h_{k-1}\right)\right] \tag{2.10}$$

onde $\left(h_0\right) = 0$ e $\left(h_n\right) = 1$, por construção. Esse coeficiente é diretamente proporcional à desigualdade da distribuição. O intervalo de valores possíveis do resultado do coeficiente de Gini é entre 0 e 1, ou seja, 0% a 100%. Um baixo valor de coeficiente de Gini indica

uma distribuição mais igualitária entre as partes, e um alto valor indica uma distribuição mais desigual. Nesta tese, trabalha-se com a hipótese de que tal valor pode ser também aplicado apropriadamente para quantificar se uma colaboração de pesquisa é igualitária em redes sociais (como será descrito na seção 3.1).

2.3 Sistemas de Recomendação

As seções anteriores apresentaram conceitos e trabalhos relacionados a redes sociais, avaliação de qualidade no contexto acadêmico e suas métricas. Completando a discussão, esta seção introduz conceitos relacionados a sistemas de recomendação, uma vez que um dos propósitos desta tese é recomendar colaboração em redes sociais acadêmicas.

Esta seção visa a apresentar alguns conceitos da área de sistemas de recomendação, incluindo os sistemas de recomendação tradicionais (seção 2.3.1) e, posteriormente, os sistemas de recomendação no contexto social (seção 2.3.2) que são um dos focos do trabalho desta tese.

2.3.1 Sistemas de Recomendação Tradicionais

A descoberta de informação digital relevante na Web é uma tarefa complexa. Sistemas de recomendação são de grande valia para atenuar os problemas associados ao fenômeno de "sobrecarga de informação", minimizando o tempo gasto para acessar as informações relevantes. Os sistemas de recomendação tradicionais surgiram nesse contexto, visando a recomendação de itens de informação. A especificação de demanda centralizada no ser humano também não é uma tarefa fácil. Pode-se experienciar essa dificuldade, por exemplo, quando se tenta buscar artigos científicos, mesmo em um bom sistema de busca e indexação como o Google Scholar[12]. A determinação de uma consulta precisa, que esteja de acordo com os requisitos do usuário, é uma tarefa que consome bastante tempo. Poucos pesquisadores têm tempo suficiente para gastar algumas horas da semana para buscar, eventualmente, novos artigos em suas áreas de pesquisa específicas. Essa funcionalidade, a especificação da consulta, pode ser realizada de diferentes formas tais como: análise do perfil do usuário, atividades, histórico, demandas por informação, etc. Nesse contexto, o desenvolvimento de sistemas de recomendação visa a ajudar os usuários a encontrarem as informações de que necessitam, de maneira pró-ativa, ou seja, sem a necessidade de eles explicitarem o que precisam.

No restante desta seção, é apresentada uma visão desses sistemas de recomendação tradicionais, incluindo as abordagens clássicas para estudo desse tipo de sistema, que são importantes para se entender o contexto de surgimento da área de pesquisa de sistemas de recomendação.

Sistemas de recomendação tradicionais provêm uma interface alternativa para tecnologias de filtragem e recuperação de informações, tendo como foco a predição daqueles itens ou partes da informação que o usuário acharia interessante e útil. Tais predições podem ser personalizadas, baseadas no perfil de cada usuário, e também podem conter julgamentos de interesses ou grau de relevância de itens previamente vistos pelo usuário. Dessa forma, sistemas de recomendação têm sido propostos e desenvolvidos para serem utilizados em diferentes contextos. Pode-se citar, por exemplo, sistemas de recomendação utilizados no auxílio à busca por informação relevante em Bibliotecas Digitais (HUANG et al., 2002; HWANG; HSIUNG; YANG, 2003; LOPES et al., 2008; SMEATON; CAL-

[12]Google Scholar: http://scholar.google.com

LAN, 2001).

Em uma perspectiva mais ampla, tais sistemas têm emergido para amenizar o problema de sobrecarga de informação na Web (BERGAMASCHI; GUERRA; LEIBA, 2010), e têm sido personalizados para diferentes domínios como livros (MOONEY; ROY, 2000), restaurantes (BURKE, 2002), filmes (GOLBECK; HENDLER, 2006), notícias (MONTANER; LÓPEZ; ROSA, 2003) e até mesmo redes sociais (MEO et al., 2011). De fato, algoritmos de recomendação têm ajudado a melhorar a qualidade de buscas na Web dentro de aplicações online tais como Amazon[13], NetFlix[14] e Google News[15]. Algoritmos de recomendação também são amplamente aplicados para ranquear sugestões em sites de comércio eletrônico, onde exploram informações sobre os compradores e definem uma lista de item recomendados (LINDEN; SMITH; YORK, 2003). Nesses sites, os sistemas de recomendação usam o conhecimento sobre o produto para guiar os consumidores na difícil tarefa de localizar produtos pelos quais eles poderão se interessar (SCHAFER; KONSTAN; RIEDL, 2001). Pode-se citar, por exemplo, os sistemas de recomendação dos sites comerciais: Amazon[16], eBay[17] e Submarino[18].

Um sistema de recomendação genérico pode ser definido como segue. Seja U o conjunto de usuários (clientes, compradores, etc.), I o conjunto de itens recomendáveis (livros, músicas, filmes, etc.) e $f(u, i)$ a função de recomendação que associa pares de (u,i) com valores orientados à aplicação (lucro, taxa, distância e afins). O objetivo de um algoritmo de recomendação é encontrar um conjunto de itens $i' \in I$ para os quais $f(u, i)$ é maximizada para o usuário.

Dada a complexidade de sistemas de recomendação, muita pesquisa tem sido desenvolvida para otimizar diferentes aspectos tais como a geração e manutenção de perfis de usuários, a técnica de filtragem, ou função de recomendação, e as conexões entre usuários. Da mesma forma, existem diferentes maneiras de classificar esses sistemas com base em tais aspectos. Por exemplo, a classificação pode considerar os agentes inteligentes empregados sobre o sistema como um todo (MONTANER; LÓPEZ; ROSA, 2003), as conexões entre usuários (PERUGINI; GONÇALVES; FOX, 2004), o tipo de entrada (BURKE, 2002) e o algoritmo de filtragem (ADOMAVICIUS; TUZHILIN, 2005; BALABANOVIC; SHOHAM, 1997). Especificamente, a última classificação define três grupos principais bem caracterizados (BALABANOVIC; SHOHAM, 1997; CLAYPOOL et al., 1999; HERLOCKER, 2000; HUANG et al., 2002): filtragem baseada em conteúdo, filtragem colaborativa e filtragem híbrida, que serão apresentados nas seções a seguir.

2.3.1.1 Filtragem baseada em conteúdo

A abordagem de filtragem baseada em conteúdo possui esse nome devido ao fato de os sistemas, que a adotam, desenvolverem a filtragem baseada em análises dos conteúdos dos itens, que possivelmente serão recomendados e podendo, também, utilizar informações do perfil do usuário. Essa abordagem trabalha com a ideia de gerar recomendações de itens relacionados ao perfil do usuário. Um perfil do item consiste de alguns atributos, que descrevam o conteúdo do item, e o perfil do usuário é criado, com base em informações, que descrevam os interesses do usuário, e relacionadas com o perfil dos itens.

[13] Amazon: http://amazon.com
[14] NetFlix: http://netflix.com
[15] Google News: http://news.google.com
[16] Amazon: http://www.amazon.com/
[17] eBay: http://www.ebay.com/
[18] Submarino: http://www.submarino.com.br/

A recomendação é gerada utilizando algumas funções de similaridade para fazer o casamento desses perfis (HUANG et al., 2002).

Para tanto, a informação precisa ser automaticamente reconhecida e categorizada, sendo gerados descritores do conteúdo de cada item. As descrições das necessidades de interesse do usuário são, ou supridas pelo usuário, como uma consulta, ou apreendidas pela observação do conteúdo dos itens consumidos pelo usuário (HERLOCKER, 2000). Então, a comparação da descrição de cada item com a descrição da necessidade de informação do usuário é utilizada para determinar se um item é ou não relevante para atender às necessidades do usuário.

Tecnologias aplicadas para filtragens baseadas em conteúdo são modelos clássicos utilizados para recuperação de informação, como: booleano, vetorial ou probabilístico (BAEZA-YATES; RIBEIRO-NETO, 2011), já que a abordagem baseada em conteúdo é derivada dos conceitos introduzidos pela comunidade de Recuperação de Informação (SHAHABI; CHEN, 2003).

A filtragem baseada em conteúdo possui algumas limitações como: o conteúdo de dados pouco estruturados é de difícil análise, por exemplo: imagens, vídeos e sons; o processamento do conteúdo do texto pode ser prejudicado devido ao uso de termos sinônimos; pode ocorrer a "super especialização", pois o sistema não recomenda itens cujo conteúdo não "case" com o perfil do usuário (REATEGUI; CAZELLA, 2005). Dessa maneira, nesse tipo de abordagem, não existe "surpresa" na recomendação, já que itens que não se relacionam com o perfil do usuário não serão recomendados ao mesmo. Além disso, segundo Claypool et al. (1999), técnicas baseadas em conteúdo têm a dificuldade de distinguir entre informação de alta e de baixa qualidade sobre o mesmo tópico. Outra limitação acontece caso o perfil do usuário seja construído a partir de informações obtidas através da sua interação com o sistema. Nesse caso, há a necessidade de o usuário ter avaliado um número suficiente de itens, antes que o sistema de recomendação possa realmente "entender" suas preferências e apresentar recomendações confiáveis. Essas recomendações serão baseadas no "casamento" entre o conteúdo dos itens a serem recomendados e o conteúdo dos itens preferidos pelo usuário (ADOMAVICIUS; TUZHILIN, 2005).

Alguns exemplos de sistemas de recomendação tradicionais baseados em conteúdo são (LANG, 1995; LIEBERMAN, 1997; LOPES et al., 2008; MAES, 1994; MOONEY; ROY, 2000).

2.3.1.2 Filtragem colaborativa

Tradicionalmente, filtragem colaborativa considera não somente o perfil do usuário, mas também informações de outros usuários para recomendar itens. Na filtragem colaborativa as ações do usuário e análises a respeito de uma informação particular são registradas para benefício de uma comunidade maior. Membros de uma comunidade podem beneficiar-se de experiências de outros, antes de consumir uma nova informação (HERLOCKER, 2000). Essa abordagem não requer nenhum tipo de descrição do conteúdo do item para que este seja recomendado. Por essa razão, a abordagem tem sido desenvolvida para cobrir áreas nas quais a filtragem baseada em conteúdo é fraca.

A filtragem colaborativa utiliza a opinião de outros usuários a respeito da informação a ser recomendada. Sistemas desse tipo podem ser não-personalizados, permitindo ao usuário descobrir itens que são de interesse popular e evitar os de desagrado popular, e podem ser personalizados, através dos relacionamentos entre perfis de usuários, trabalhando com a ideia de que, se os interesses dos mesmos são similares, itens preferidos por um podem

ser recomendados a outros com perfil similar ou à comunidade que este usuário faz parte. Tais relacionamentos entre usuários podem ser informados ao sistema ou descobertos de forma automática, com base na análise de padrões comuns nas avaliações dos itens.

Nesse tipo de abordagem, podem ocorrer problemas como a "partida fria" quando não estão inicialmente disponíveis dados sobre o perfil do usuário, não havendo informações que possibilitem encontrar um perfil similar. Ou ainda, segundo Balabanovic e Shoham (1997), se um novo item for adicionado na base de dados, não existe meio de este ser recomendado até que um usuário o avalie ou especifique outro item já avaliado como similar a este. Quando há um número de usuários relativamente pequeno para o volume de informação do sistema, existe o risco de a cobertura das avaliações dos itens tornar-se muito esparsa, diminuindo a coleção de itens recomendáveis. Mais ainda, recomendações de itens recentes na base de dados podem ser inexatas, porque existem poucas avaliações para basear as predições da filtragem colaborativa. Além disso, segundo Claypool et al. (1999), em pequenas ou até médias comunidades de usuários, existem indivíduos que não se beneficiam de sistemas de filtragem colaborativa puros, porque suas opiniões não estão consistentemente de acordo ou em desacordo com qualquer grupo de pessoas, são conhecidos como usuários com gostos incomuns.

Alguns exemplos de sistemas de recomendação tradicionais de filtragem colaborativa são Tapestry (GOLDBERG et al., 1992) , GroupLens/MovieLens (KONSTAN et al., 1997; RESNICK et al., 1994) e Ringo/Firefly (SHARDANAND; MAES, 1995).

2.3.1.3 Filtragem híbrida

As abordagens, de filtragem baseada em conteúdo e filtragem colaborativa, não são mutuamente exclusivas, existindo inúmeros esforços para integração de ambas, a fim de obter maior exatidão nas recomendações (HUANG et al., 2002). Para tanto, a abordagem de filtragem híbrida surge como uma combinação dessas duas abordagens (apresentadas nas seções 2.3.1.1 e 2.3.1.2), buscando agregar as características de cada uma delas e solucionar as limitações encontradas, de forma a melhor atender às necessidades dos usuários. Com o uso de tal combinação, podem ser alcançados os benefícios da filtragem baseada em conteúdo, que inclui a predição para todos os itens e usuários, sem a dependência do número de usuários e do número de avaliações dos itens, enquanto se ganha em exatidão nas predições de filtragem colaborativa conforme o número de usuários e avaliações cresce (CLAYPOOL et al., 1999). Esse tipo de função de recomendação considera toda a informação disponível para construir um resultado: a descrição dos itens, o perfil dos usuários e as informações de outros usuários.

Segundo Reategui e Cazella (2005), algumas das características importantes, herdadas pela filtragem híbrida de cada uma das abordagens, podem ser observadas na Figura 2.7, são elas: (i) descoberta de relacionamentos entre usuários; (ii) recomendação baseada na avaliação de itens; (iii) trato de usuários incomuns; (iv) precisão independente do número de usuários e avaliações.

As características (i) e (ii) são herdadas da filtragem colaborativa, já que esta trabalha com a ideia de "perfis similares". Na filtragem baseada em conteúdo não é levado em consideração qualquer tipo de relacionamento entre perfis de usuários. Além disso, itens com histórico de boa recepção por diversos tipos de usuários também não são relevantes na filtragem baseada em conteúdo, pois esta não gera a recomendação de itens não relacionados ao perfil do usuário (sem "surpresa" na recomendação).

As características (iii) e (iv) são alcançadas graças à abordagem baseada em conteúdo. Se fosse utilizada somente a filtragem colaborativa, não seria possível obter bons resul-

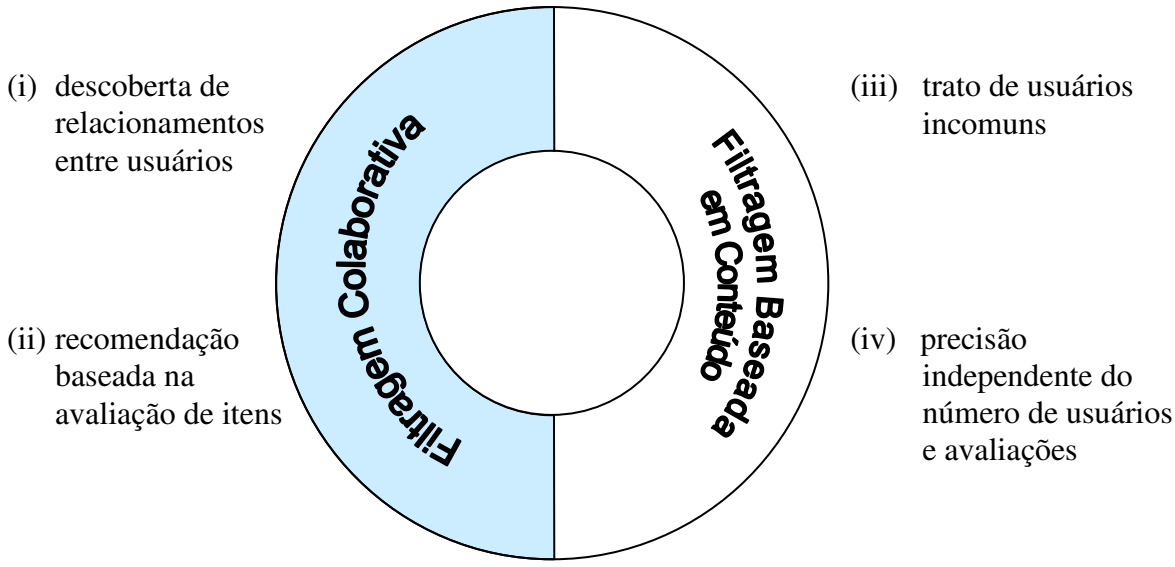


Figura 2.7: Características herdadas pela Filtragem Híbrida (adaptado de (REATEGUI; CAZELLA, 2005)).

tados para usuários incomuns, pois não se conseguiria um perfil de usuário semelhante para "casar" com o perfil destes usuários, assim como, havendo poucos usuários surge dificuldade na obtenção de informações para casamento entre os perfis. Isso já não ocorre na filtragem baseada em conteúdo.

Segundo Huang et al. (2002), sistemas híbridos podem obter diferentes graus de ganho em exatidão de predição, por utilizarem múltiplas fontes de informação, variando de modestos benefícios a melhorias significativas. Porém, essa adição de informação nem sempre conduz a melhores resultados. A análise da variação na qualidade da recomendação, em função da multidimensionalidade da informação, requer um estudo bastante aprofundado.

Alguns exemplos de sistemas de recomendação tradicionais híbridos são Fab (BALABANOVIC; SHOHAM, 1997), P-Tango (CLAYPOOL et al., 1999) e (PAZZANI, 1999).

2.3.2 Sistemas de Recomendação Social

Como comentado anteriormente, a popularização de redes de relacionamento (redes sociais online) foi uma das responsáveis pelo crescimento de pesquisas e estudos sobre Redes Sociais. Nesse cenário, surgiram abordagens sendo desenvolvidas para sistemas de recomendação considerando o contexto social. Além disso, o foco de recomendação, que anteriormente era exclusivamente de determinados itens de informação, passou também a vislumbrar a recomendação de colaborações entre os próprios usuários da rede social analisada no processo de geração de recomendações.

Nesse contexto, pode-se citar, por exemplo, os sistemas de recomendação de possíveis "amigos" que são apresentados em sites de redes de relacionamento como Facebook[19] e Orkut[20]. Nesses, a recomendação é realizada com base principalmente nas relações sociais, de amizade, estabelecidas entre os indivíduos, visando recomendar aqueles que estão em proximidade social através da análise de conexões de "amigos" de "amigos".

Para determinação dessa proximidade social entre indivíduos, muitas vezes, são uti-

[19]Facebook: http://www.facebook.com
[20]Orkut: http://www.orkut.com

lizadas métricas originárias da teoria dos grafos voltadas para as redes sociais (LIBEN-NOWELL; KLEINBERG, 2007; NEWMAN, 2003; QUERCIA; CAPRA, 2009). Normalmente, os métodos atribuem um escore (valor de peso), nomeado $score(x, y)$, para todos os pares de nós $< x, y >$ pertencentes ao grafo da rede social sob análise, denominado G. Com os resultados de escore para todos os pares da rede social ordenados decrescentemente, tem-se uma lista que indica os *links* "previstos" em ordem decrescente de confiança.

Um método tradicional e bastante utilizado é o do menor caminho (*shortest path*) (CHERKASSKY; GOLDBERG; RADZIK, 1996), onde o $score(x, y)$ é calculado pelo tamanho do menor caminho que une os nós x e y dentro do grafo G sendo considerado. Então, os valores de escore serão ordenados crescentemente, já que quanto menor o caminho maior será a confiança do link previsto. Tal abordagem utiliza a noção de mundo pequeno, em que todos os indivíduos da rede estão conectados por caminhos curtos. Outro método utilizado é o dos vizinhos comuns (*common neighbors*) que calcula o $score(x, y)$ de acordo com a vizinhança de um determinado nó. Nessa abordagem, quanto maior a quantidade de vizinhos compartilhados pelos nós x e y, maior será a proximidade desses nós dentro da rede social representada pelo grafo G. Ainda outro exemplo de abordagem é a métrica de Katz (1953), nessa, o $score(x, y)$ é calculado como sendo uma soma ponderada dentro de todos os caminhos existentes entre dois nós x e y. Liben-Nowell e Kleinberg (2007) apresentam um trabalho sobre preditores baseados em diferentes métricas de proximidade de grafos.

No restante desta seção, são apresentadas novas perspectivas no estudo de sistemas de recomendação com a consideração do contexto social e algumas de suas particularidades (seção 2.3.2.1) e, por fim, são comentados alguns trabalhos relacionados em recomendação social (seção 2.3.2.2).

2.3.2.1 Novas perspectivas de estudo

Atualmente, mudanças nas próprias abordagens sendo desenvolvidas para novos sistemas de recomendação, considerando o contexto social, levaram à necessidade de novas formas de se estudar a área de sistemas de recomendação. Surgiram os sistemas que ficaram também conhecidos como sistemas de "casamento" social que visam à recomendação de pessoas umas às outras, ao invés de recomendar itens de informação. A recomendação de pessoas difere da recomendação de itens. Algumas questões são levantadas quando se trata desse tipo de recomendação, pois, para recomendar pessoas, algumas informações pessoais estão envolvidas e precisam ser necessariamente reveladas. Com isso, surgem alguns riscos de interação inerentes, levantando questões como: privacidade, confiança, reputação e atração interpessoal (TERVEEN; MCDONALD, 2005).

Terveen e McDonald (2005) propuseram um modelo básico para ilustrar o processo de "casamento" social, apresentado na Figura 2.8. Nesse modelo, o casamento social é representado constituindo-se de quatro passos principais. O primeiro consiste da *Modelagem* do conjunto de usuários a serem "casados", ou seja, é feita a construção do perfil dos usuários. O segundo inclui o *Casamento* de usuários e sua recomendação em resposta a requisições explícitas ou oportunidades implícitas, para tanto, algum algoritmo de "casamento" deve ser aplicado. O terceiro é a *Introdução* de usuários "casados". O quarto inclui a *Interação* que pode ocorrer entre os usuários, ou através de um espaço mediado, criado pelo próprio sistema, ou através de outros meios da própria escolha dos usuários. Os usuários, a qualquer momento, poderão atualizar seus perfis, se não estiverem satisfeitos com o tipo de pessoas que estão sendo recomendadas.

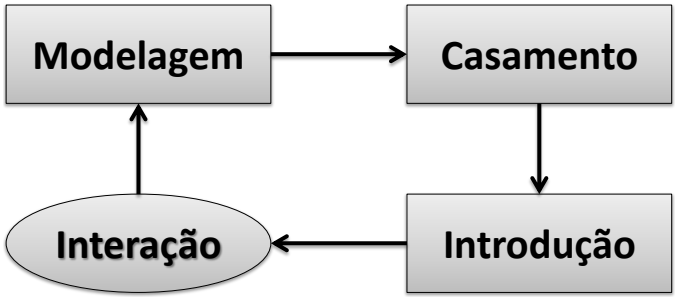


Figura 2.8: Modelo do processo de "casamento" social (adaptado de (TERVEEN; MCDONALD, 2005)).

Além disso, Terveen e McDonald (2005) também levantaram algumas questões importantes a serem consideradas nos sistemas de "casamento" social:

- *Modelagem do perfil do usuário*. Que tipo de informação um sistema representa sobre seus usuários, e como esta informação é adquirida?
- *Computação dos "casamentos"*. Qual é o modelo de sistema que faz bons casamentos? Como o sistema computa esses casamentos?
- *Introdução*. Como as pessoas "casadas" são reunidas? Que tipo de informação o sistema revela sobre a pessoa?
- *Interação*. Em que medida o sistema facilita a interação? A interação acontece em um espaço mediado, criado pelo sistema, ou os usuários interagem como eles bem entendem, inclusive ao vivo?
- *Feedback*. Como os resultados de uma interação refletem nos perfis dos usuários? O sistema pode automaticamente atualizar o perfil do usuário, ou o usuário deve prover um *feedback* explícito?

Conforme destacam Perugini, Gonçalves e Fox (2004), sistemas de recomendação estão inseridos em um contexto social, uma vez que as recomendações são entregues a um usuário ou a uma comunidade de usuários. Esse contexto social envolve, mesmo que informalmente, comunidades de usuários, que podem ser vistas como redes sociais. Dessa forma, Perugini, Gonçalves e Fox (2004) propõem uma abordagem de análise centrada nas conexões (*connection-centric approach*) para o estudo de sistemas de recomendação.

Perugini, Gonçalves e Fox (2004) enfatizam que a recomendação tem um inerente elemento social e, em última análise, destina-se a conectar pessoas, quer diretamente como resultado da modelagem explícita do usuário, ou indiretamente através da descoberta de relacionamentos implícitos nos dados existentes. A modelagem dos usuários pode ser vista como a base para computar sobreposições de interesses e conduzir a identificação de "conexões" entre pessoas para gerar as recomendações.

Na abordagem centrada nas conexões, os sistemas são caracterizados pela forma como obtêm as informações sobre os usuários que utilizam para o estabelecimento das conexões entre estes: explicitamente ou implicitamente. Na Figura 2.9, é apresentado um esquema da visão centrada nas conexões da recomendação como "conectora" de indivíduos em uma rede social. Esta rede social pode ser formada pela coleta explícita de avaliações ou perfis do usuário; ou pela identificação e descoberta de uma rede pela exposição de

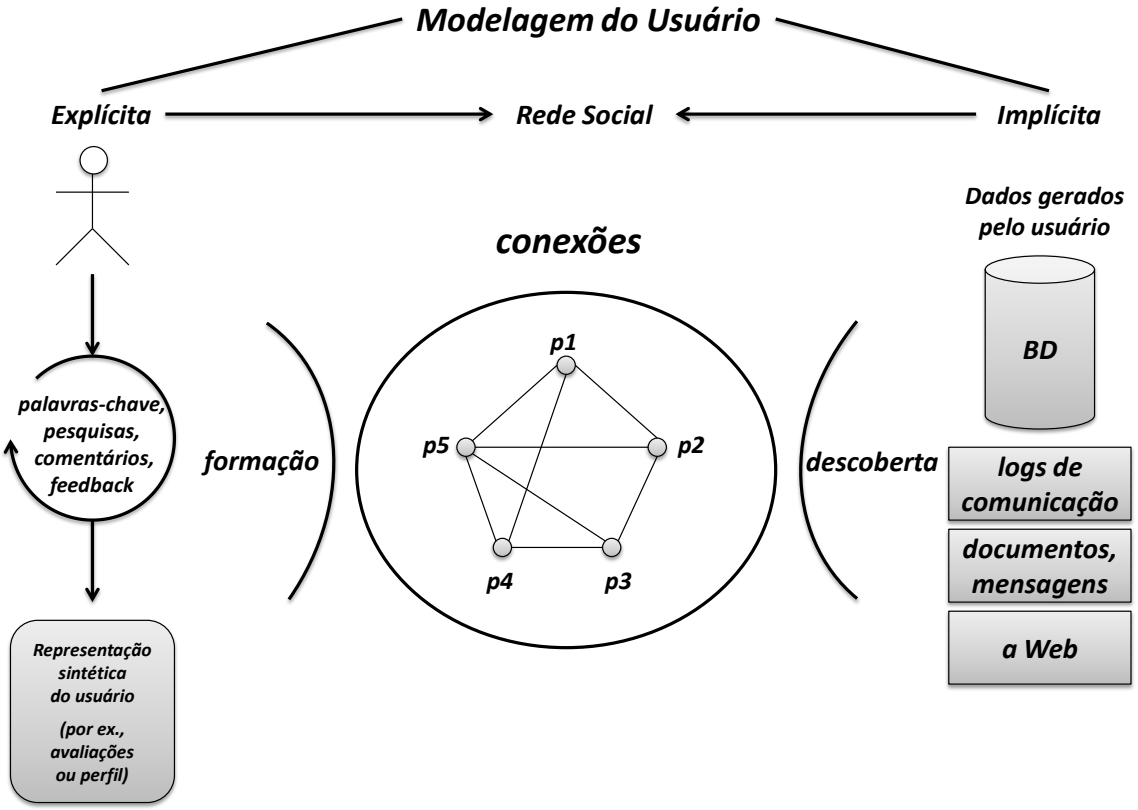


Figura 2.9: Esquemático da visão centrada nas conexões (adaptado de (PERUGINI; GONÇALVES; FOX, 2004)).

comunidades implícitas propriamente organizadas em dados gerados pelo usuário, tais como comunicação ou *logs* da web. Embora não ilustrado explicitamente nessa figura, essas duas abordagens também podem ser combinadas.

A grande mudança em relação às abordagens anteriormente existentes em sistemas de recomendação tradicionais consiste principalmente na classificação centrada em conexões em que os dados são obtidos implicitamente. Os dados obtidos de forma explícita, para construção de uma possível rede social, já se refletiam nas abordagens colaborativas em que a avaliação dos itens era utilizada para obtenção de usuários com comportamento semelhante nessa avaliação. Não precisar dessa avaliação explícita dos usuários dos itens recomendados e ainda assim utilizar relações entre esses, para obtenção de recomendações, é um dos grandes diferenciais das novas abordagens em sistemas de recomendação. Essa nova visão trouxe avanços e desenvolvimentos originalmente da área de redes sociais sendo aplicados na construção de novas abordagens de sistemas de recomendação que obtêm os dados, de forma implícita, para formação de redes sociais.

No contexto de web social, outra visão interessante para analisar sistemas de recomendação é sobre o tipo de propriedades que estão sendo consideradas para a construção do método de recomendação: (i) *propriedades estruturais* que consideram a análise das relações entre os atores de uma rede social e (ii) *propriedades semânticas* que consideram a relação dessas propriedades estruturais com o contexto semântico. As propriedades estruturais são utilizadas pela maioria das abordagens que se originam de métodos tradicionais utilizados em análises de redes sociais. Por outro lado, as propriedades semânticas nem sempre são consideradas nessas abordagens. Figueira Filho, Albuquerque e Geus (2008) destacam a tendência de que as novas propostas em sistemas de recomendação unifiquem

propriedades estruturais com propriedades semânticas.

2.3.2.2 Abordagens relacionadas

Nesta seção, serão apresentados alguns trabalhos relacionados a Sistemas de Recomendação Social. Como um dos objetivos desta tese é a recomendação de colaborações acadêmicas, alguns desses trabalhos são diretamente relacionados a sistemas de recomendação no contexto de redes sociais acadêmicas ou podem ser aplicados nesse contexto. Além disso, alguns deles são associados à Web Semântica e enfatizam questões relacionadas (por exemplo, ontologias).

A Web Semântica (*Semantic Web*), idealizada por Berners-Lee e Fischetti (1999), é uma extensão da Web atual, na qual a informação é gerada, não somente para leitores humanos, mas também para processamento por máquinas, possibilitando serviços de informação inteligentes, sites Web personalizados e máquinas de busca semanticamente enriquecidas. Para atingir esta meta, um dos importantes requisitos consiste no desenvolvimento de ontologias para criação de um modelo de nova geração da Internet.

As estruturas conceituais que definem uma ontologia provêm uma chave para dados processáveis por máquinas na Web Semântica. Ontologias servem como um esquema de metadados, provendo um vocabulário controlado de conceitos, cada qual definido explicitamente e semanticamente processável por máquina. Pela definição de teorias de domínio comum e compartilhado, ontologias ajudam pessoas e máquinas a se comunicarem concisamente - suportando intercâmbio de semântica, não apenas de sintaxe. Por isso, o sucesso e a proliferação da Web Semântica dependem da construção de ontologias de domínio específicas (MAEDCHE; STAAB, 2001).

Também se buscou como foco os trabalhos que recomendam colaborações entre usuários ou indicam algum *expert* em determinado contexto. Os sistemas de descoberta de especialistas (*Expert Finding Systems* - EFS) possibilitam que os usuários descubram especialistas (*experts*) em um determinado assunto a fim de contatar ou adquirir seus conhecimentos.

Para considerar os aspectos de interatividade de Redes Sociais, o restante desta seção descreve trabalhos relacionados que podem ser classificados como sistemas de recomendação colaborativos e híbridos. Muito embora abordagens puramente baseadas em conteúdo possam ser usadas para recomendar indivíduos (por exemplo, algoritmo *Content Matching* de Chen et al. (2009) faz recomendação de pessoas com perfis similares em conteúdo), está-se interessado em considerar a informação que pode ser extraída da interação entre indivíduos modelada por uma rede social, e não apenas aspectos de conteúdo relacionados aos itens que serão recomendados. Especificamente, muitas abordagens utilizam o contexto social para melhorar a recomendação de itens como filmes (GOLBECK; HENDLER, 2006), artigos (HWANG; WEI; LIAO, 2010; WENG; CHANG, 2008; ZAIANE; CHEN; GOEBEL, 2007; ZANARDI; CAPRA, 2008), mídia social (GUY et al., 2010) e até mesmo consultas em mecanismos de busca (LI; OTSUKA; KITSUREGAWA, 2010). Estas abordagens utilizam informações da rede social para inferir possíveis itens de interesse, empregando as relações entre indivíduos. Por exemplo, relacionamentos entre usuários com gostos em comum, itens relacionados a "amigos" do usuário alvo de recomendação, itens de indivíduos relacionados a outros itens que já interessaram ao usuário alvo, entre outros.

Abordagens correlatas de Filtragem Colaborativa. A informação mais utilizada de uma rede social para recomendação é provavelmente suas conexões (por exemplo, os

amigos de seus amigos). Um grafo de conexões pode ser empregado, ou combinado, para uma variedade de propósitos (como encontrar novos amigos). Por exemplo, Quercia e Capra (2009) propõem um arcabouço (*framework*), chamado *FriendSensing*, para automaticamente sugerir amigos para usuários de uma rede social baseada na proximidade física entre eles, a qual é identificada através da localização de dispositivos móveis. A rede social entre os usuários é construída com pesos atribuídos às relações de acordo com a duração e a frequência dos "encontros" entre usuários. Com base em algoritmos de proximidade geográfica e predição de links, é gerada como recomendação uma lista de pessoas que o usuário pode conhecer. Do mesmo modo, Chen et al. (2009) propõem dois algoritmos (chamados FoF e SONAR) para recomendar pessoas na rede social Beehive da IBM. O algoritmo FoF (*Friend-of-Friend*) considera somente as informações estruturais da rede social, baseando-se na recomendação de "amigos" de "amigos". O algoritmo SONAR, é baseado no sistema SONAR que agrega informações de relacionamento social de diferentes fontes de dados públicas dentro da IBM, como por exemplo: estrutura organizacional, base de dados de publicações, base de patentes, etc. Para cada uma das informações, o sistema computa um escore de relacionamento normalizado entre duas pessoas. Esses escores são agregados para unificar um único escore que será utilizado para ranquear a recomendação de pessoas.

Utilizando diferentes perspectivas, Ogata et al. (2001) propõem o sistema *PeCo-Mediator-II* para buscar especialistas (*experts*) através de uma cadeia de conexões pessoais, dada pela troca de emails e cartões pessoais. A função de recomendação analisa o histórico (*logs*) de interação dentro da rede, a fim de encontrar um especialista adequado e recomendá-lo de volta ao usuário. Da mesma maneira, Karagiannis e Vojnovic (2009) constroem uma rede social através das interações de troca de emails dentre um grupo de pessoas e recomenda novos "amigos de amigos" (contatos dos contatos).

Abordagens correlatas de Filtragem Híbrida. Algumas abordagens consideram tanto a análise de questões estruturais de uma rede social quanto de conteúdo para gerar as recomendações. Por exemplo, o sistema *ReferralWeb* (KAUTZ; SELMAN; SHAH, 1997) identifica especialistas e gera um caminho de conexão entre o usuário alvo de recomendação e o especialista a ser recomendado, com base em uma rede social, a qual é construída pela mineração de dados públicos disponibilizados em documentos Web. Também, McDonald (2003) detalha uma avaliação de diferentes redes sociais que podem ser usadas para recomendar especialistas. Como exemplo de um sistema mais complexo, *FilmTrust* (GOLBECK; HENDLER, 2006) é um site Web que integra análises de Web Semântica para uma rede social com o conceito de confiança (*trust*), a fim de gerar a recomendação de filmes. Usuários provêm uma taxa de confiança para cada uma de suas conexões, e itens (filmes) são avaliados (revisados) pelos usuários. Então, uma função de recomendação agrega ambos os dados (das conexões dos usuários e das avaliações dos itens), computa uma classificação geral para todos os itens e apresenta uma lista ordenada como recomendação ao usuário. A ideia básica é que maiores taxas de confiança podem ser associadas a usuários que sugerem filmes com uma maior acurácia. A rede social irá indicar as taxas de confiança entre usuários pela consideração dos relacionamentos entre usuários. Além disso, Chen et al. (2009) propõem o algoritmo *Content-plus-Link* (CplusL), que utiliza o casamento baseado em conteúdo e acrescenta uma informação de "link" social derivada da estrutura da rede social sob análise. Nessa abordagem, "link" social válido é definido como um grau de separação entre usuários de no máximo três ou quatro usuários. O objetivo do algoritmo é favorecer candidatos à recomendação pela abordagem baseada em conteúdo que tiverem links válidos.

Redes Sociais e Análise Acadêmica. Análises sociais e academia têm caminhado juntas. Diferentes comunidades de pesquisa têm explorado os benefícios da análise de redes sociais para entender suas próprias características e comportamentos (DING, 2011; WANG et al., 2010). Exemplos incluem áreas com a física (NEWMAN, 2001), matemática (BARABÁSI, 2002), banco de dados (NASCIMENTO; SANDER; POUND, 2003), bibliotecas digitais (LIU et al., 2005) e recuperação de informações (SMEATON et al., 2003). Com os avanços de redes sociais orientadas à academia, tais como ArnetMiner[21], Microsoft Academic Search[22], LinkedIn[23] e ResearchGate[24], diferentes serviços podem ser providos buscando atender às necessidades emergentes nesse contexto, tais como:

- Detecção de Conflito de Interesse. Aleman-Meza et al. (2006) constrõem uma abordagem com base em duas redes sociais, construídas utilizando dados da DBLP e FOAF (*Friend-Of-A-Friend*), a fim de detectar relacionamentos de conflito de interesse entre autores de artigos científicos e potenciais revisores de seus artigos. Algumas abordagens e métricas para construção da rede e determinação dos pesos dos relacionamentos entre autores são propostas. Além disso, regras são estabelecidas para determinar um possível grau de conflito de interesse, com base na rede social construída e nos pesos dos relacionamentos entre autores.

- Usuários relacionados. Weng e Chang (2008) propõem um método de recomendação que utiliza ontologias e o modelo de ativação espalhada para buscar por outros usuários influentes em uma rede social acadêmica. Os trabalhos anteriormente citados *Peco-Mediator-II* (OGATA et al., 2001) e *ReferralWeb* (KAUTZ; SELMAN; SHAH, 1997) também buscam por especialistas em uma rede social interpessoal, que pode então ser especializada para um cenário acadêmico.

- Colaborações. *DBConnect* (ZAIANE; CHEN; GOEBEL, 2007) explora uma rede social codificada na base de dados da DBLP, a fim de revelar conhecimento interessante sobre a comunidade de pesquisa e eventualmente recomendar colaborações. Essa abordagem também ajuda o usuário a buscar conferências relevantes, autores similares e tópicos de pesquisa interessantes. A recomendação de colaboradores é baseada em um grafo tripartido (autor-conferência-tópico) e objetiva recomendar autores com tópicos e experiências em conferências similares (ZAÏANE; CHEN; GOEBEL, 2009).

- Recomendação de Literatura. Hwang, Wei e Liao (2010) propõem um método que constrói uma rede de coautoria e recomenda artigos considerando o perfil de tarefas do usuário e tal rede.

2.4 Enquadramento desta tese em comparação aos trabalhos relacionados

Nas seções anteriores deste capítulo, foi apresentada uma visão geral das áreas de estudo em que esta tese está inserida e discutiu-se uma série de trabalhos relacionados. Por fim, esta seção, mostra o enquadramento desta tese em comparativo aos trabalhos relacionados, sendo reforçadas as novidades de suas contribuições (previamente discutidas no

[21] ArnetMiner: http://www.arnetminer.com
[22] Microsoft Academic Search: http://academic.research.microsoft.com
[23] LinkedIn: http://www.linkedin.com/
[24] ResearchGate: http://www.researchgate.net

Capítulo 1). Para uma melhor organização desta seção, ela é subdividida nas duas grandes áreas de investigação principais desta tese: análise e avaliação de grupos de pesquisadores (seção 2.4.1) e recomendação de colaborações acadêmicas (seção 2.4.2).

2.4.1 Análise e Avaliação de grupos de pesquisadores

O trabalho desta tese está inserido no contexto de avaliação de qualidade na academia. Entretanto, ao contrário dos trabalhos relacionados anteriormente (apresentados na seção 2.2.1), o presente trabalho não considera estatísticas de citações. A novidade está justamente na visão de explorar outra faceta para inferir qualidade. A nova faceta, nesse caso, é a consideração de um aspecto social para avaliar grupos de pesquisadores, através de suas interações em colaborações. Especificamente, foram definidos indicadores de qualidade para avaliação de grupos de pesquisa relativos à análise da colaboração interna dos seus pesquisadores membros. Portanto, diferentemente dos trabalhos prévios, são introduzidas métricas de análise de redes sociais para estimar tais indicadores. A maioria dos trabalhos relacionados em redes sociais utiliza análises com propósitos comparativos e para entender o comportamento das interações, mas eles não objetivam inferir qualidade ou construir um ranking de grupos de pesquisa como é o caso desta tese.

A presente tese apresenta um comparativo entre diferentes métricas de análise de redes sociais para determinação de qualidade no contexto acadêmico e ainda propõe outras para avaliação de redes sociais. Essas novas métricas se mostraram mais adequadas para avaliar a colaboração interna entre pesquisadores e, por consequência, para mensurar indicadores definidos para avaliar a qualidade de grupos de pesquisa. O trabalho desta tese mostra que a qualidade de grupos de pesquisa está associada ao nível de colaboração interna e à forma de distribuição dessas colaborações.

Um dos únicos trabalhos encontrados na literatura que faz uso de análises de redes sociais com objetivos de ranqueamento em redes sociais acadêmicas é o de Freire e Figueiredo (2010). Entretanto, os seus autores determinam uma métrica para ranqueamento de vértices e grupos de vértices, através da intensidade dos relacionamentos considerando cortes e o peso do vértice no corte do grafo representando uma rede de coautoria. O indicador de qualidade focado por esse trabalho é relacionado às ligações externas quando é efetuado um corte na rede - para considerar apenas o(s) pesquisador(es) desejado(s). Os resultados para ranqueamento de indivíduos são destacados, principalmente para identificar os pesquisadores mais influentes. Em comparativo, outro ponto a destacar é que no trabalho desta tese focaliza-se principalmente na avaliação de grupos e utiliza-se um outro conjunto de análise (colaborações internas).

Especificamente sobre análises utilizando o coeficiente de Gini, alguns trabalhos relacionados são comentados a seguir e comparados com a proposta da presente tese. Sobre os atuais usos do coeficiente de Gini, os autores Pissard e Prieur (2007) propuseram técnicas para quantificar parâmetros em redes sociais. Eles consideram o contexto de usuários em uma rede social sobre o *upload* de fotos. Nesta tese, considera-se um cenário completamente diferente, em que se deseja quantificar a colaboração científica em redes sociais de pesquisa. Em outro trabalho relacionado, Silva et al. (2010) aplicam o coeficiente de Gini para analisar distribuições de índice h para ranquear comitês de programa de conferências e conselhos editoriais de periódicos. No entanto, os autores não aplicam o coeficiente para analisar as conexões de redes sociais, mas para analisar a distribuição dos índices h, explorando a faceta bibliométrica de citações. Nesta tese, diferentemente dos trabalhos anteriores, o coeficiente de Gini é aplicado para analisar redes sociais de pesquisa. Além disso, tal coeficiente é empregado em duas diferentes perspectivas: para analisar como

uma rede evolui em termos de produtividade e distribuição de conectividade (seção 3.3.1) e para avaliar grupos de pesquisa (seção 3.3.2).

2.4.2 Recomendação de colaborações no contexto acadêmico

O trabalho desta tese focaliza na recomendação de colaborações em redes sociais acadêmicas. Nessa direção, a maioria dos sistemas relacionados objetiva recomendar especialistas e novas colaborações em redes de outros contextos (apresentados na seção 2.3.2.2). Por exemplo, muitas abordagens surgiram para recomendação de possíveis amigos, tais como (CHEN et al., 2009; KARAGIANNIS; VOJNOVIC, 2009; QUERCIA; CAPRA, 2009). Em tais abordagens, os sistemas definem recomendações com base em relações sociais (de amizade) entre usuários por proximidade social (CHEN et al., 2009), trocas de emails (KARAGIANNIS; VOJNOVIC, 2009) e proximidade física (QUERCIA; CAPRA, 2009). Muito embora a abordagem de recomendação apresentada nesta tese compartilhe o objetivo de recomendar pessoas, ela foi adaptada para trabalhar no cenário acadêmico. Nesse sentido, as abordagens prévias que estão mais próximas da abordagem desta tese incluem aquelas para recomendar especialistas (KAUTZ; SELMAN; SHAH, 1997; MCDONALD, 2003; OGATA et al., 2001). No entanto, elas não consideram a área de pesquisa do usuário alvo (que receberá a recomendação).

No trabalho desta tese, a abordagem de recomendação considera as áreas de pesquisa do usuário alvo e tenta "casá-las" com usuários que possuam perfis similares. Por isso, ela considera ambos, os aspectos baseados em conteúdo (para encontrar pesquisadores das mesmas áreas de pesquisa) e aspectos estruturais (para análise de relacionamentos existentes e a determinação de proximidade social entre pesquisadores). A consideração do perfil do usuário alvo de recomendação não era considerada pelas abordagens anteriores (em recomendação de especialistas) e apenas recentemente passou a ser explorada. Além disso, alguns trabalhos recentes, desenvolvidos paralelamente ao trabalho desta tese, surgiram propondo soluções relacionadas à análise e recomendação de colaborações no contexto acadêmico. Isso mostra a importância e atualidade do tema abordado. A seguir, são comentados alguns desses trabalhos mais diretamente relacionados a esta tese.

Uma abordagem, das poucas encontradas, que recomenda colaborações no contexto acadêmico e que considera o perfil do usuário alvo de recomendação é o trabalho de Zaïane, Chen e Goebel (2009). Tal trabalho explora similaridades entre tópicos e experiências em conferências entre pesquisadores para propor uma abordagem de recomendação de colaboradores. Entretanto, este artigo não apresenta qualquer avaliação de resultados. Uma abordagem similar é apresentada em (XU et al., 2010). Essa última trabalha com uma rede de duas camadas (rede social heterogênea em relação aos nós e arestas) para gerar recomendações através de algoritmos de caminhamento randômico. Nesta rede, uma camada modela os pesquisadores e suas relações sociais, não especificadas claramente no artigo, outra camada modela termos para indicar as especialidades (conceitos da WordNet[25]) e suas relações de intersecção, e, a ligação entre as duas camadas é feita pela associação entre pesquisadores e termos. Já o trabalho de Giuliani, De Petris e Nico (2010) explora a avaliação de colaborações através de relações e compartilhamento de recursos. Nesse, a proposta é estimar o potencial de colaboração de um grupo e avaliar se o grupo está ou não aproveitando esse potencial identificado, não objetiva recomendação. Tal proposta é realizada pelo estudo dos conjuntos de artigos compartilhados e não

[25] WordNet: http://wordnet.princeton.edu

compartilhados, entre pesquisadores, que contenham os mesmos termos. Em (HECK; HANRATHS; STOCK, 2011), são exploradas duas facetas relativas a citações, para recomendação de especialistas. A ideia é de que sejam recomendados, através de métodos de filtragem colaborativa, pesquisadores que citaram as mesmas referências em suas publicações prévias e/ou que tenham tido suas publicações prévias citadas conjuntamente em outros artigos. Além desses, o trabalho de Chen et al. (2011), propõe uma ferramenta de busca para descoberta de colaborações, nomeada CollabSeer. A CollabSeer retorna recomendações de colaboradores, com base na estrutura da rede de coautorias, entre usuário alvo e possíveis recomendáveis, e pode ser refinada, para retornar recomendações mais acuradas pela seleção, feita explicitamente por parte do usuário, de tópicos de interesse desejados.

Além disso, é muito importante entender que a adaptação de um sistema de recomendação para o cenário acadêmico não é uma tarefa trivial. Por exemplo, para sugerir novos amigos em uma rede social comum, um número de amigos em comum pode ser preponderante para estimar a proximidade social entre usuários. Entretanto, em um contexto acadêmico, proximidade social tem uma interpretação distinta. Esta deve considerar não somente a conexão social entre pessoas, mas também sua formação acadêmica. Tal informação pode ser dada por conexões em trabalhos prévios, artigos em coautoria, áreas de pesquisa em comum e muitos outros. Além disso, o tipo de recomendação pode ser diferenciado. Especificamente, esta tese considera essa visão e coloca-a em ação pela recomendação não apenas de novas colaborações, mas também da intensificação de colaborações já existentes. No geral, podemos dizer que o objetivo final da abordagem desta tese é mais amplo do que o dos trabalhos anteriores aqui descritos, uma vez que estimula novas colaborações, bem como as já existentes, que são muito importantes no contexto acadêmico.

Outro ponto interessante explorado por esta tese é a consideração de aspectos temporais relativos à ocorrência das colaborações acadêmicas consideradas na geração das recomendações. Nesse sentido, a seguir, são retomados alguns trabalhos relacionados em Sistemas de Recomendação Social e são enfatizados trabalhos em redes sociais, que abordam questões temporais. Muitas abordagens em recomendação que utilizam redes sociais consideram alguns aspectos estruturais da rede para gerar recomendações. Para estabelecer escores que estimam a proximidade entre os atores, modelados por uma Rede Social, e até mesmo para fazer a predição de novas ligações, muitos trabalhos empregam métricas de teoria de grafos (LIBEN-NOWELL; KLEINBERG, 2007; NEWMAN, 2003; QUERCIA; CAPRA, 2009). Liben-Nowell e Kleinberg (2007) apresentam um estudo detalhado sobre o uso de diferentes métodos para predição de ligações em redes sociais. Entretanto, o trabalho desses autores não explora métodos de estabelecimento de pesos para as ligações e nem o impacto destes nos resultados das funções de escore.

Dessa forma, pode-se observar que em redes sociais acadêmicas (redes de coautoria), a rede pode ser sem a definição dos pesos das ligações (KAUTZ; SELMAN; SHAH, 1997; LIBEN-NOWELL; KLEINBERG, 2007; MCDONALD, 2003). Por outro lado, para uma rede com a definição dos pesos das ligações, o coeficiente de *Jaccard* é usualmente aplicado (ALEMAN-MEZA et al., 2006; HWANG; WEI; LIAO, 2010). No trabalho desta tese, vai-se um passo adiante, e explora-se a influência dos aspectos temporais nos pesos das ligações que são utilizados para determinar os resultados das recomendações.

Recentes propostas têm abordado a importância de considerar aspectos temporais na recomendação de itens dentro de um contexto social. Por exemplo, Hwang, Wei e Liao (2010) propõem uma abordagem focada em tarefa para recomendar literatura e que consi-

dera informação temporal, através de *logs* de uso, para compor um perfil focado em tarefa. Consequentemente, os aspectos temporais são tratados no nível da geração do perfil do usuário. Para análises de redes sociais, não com propósitos de recomendação, Tang et al. (2009) propõem novas métricas de distância temporal para quantificar e comparar a velocidade (*delay*) de processos de difusão de informação, considerando a evolução de uma rede. Xiang et al. (2010) apresentam uma abordagem de recomendação temporal considerando preferências dos usuários de curta e longa duração. Esta abordagem lida com as mudanças de preferências ao longo do tempo. Entretanto, o foco de tal recomendação são itens, não indivíduos ou colaboradores. Dessa forma, por não terem sido encontrados outros trabalhos relacionados, o trabalho desta tese é um dos primeiros a estudar a influência de aspectos temporais (através da ponderação das relações na SN considerando este aspecto) na recomendação de colaboradores para Redes Sociais Acadêmicas.

Neste capítulo foi apresentada uma visão geral sobre os principais temas de pesquisa relacionados com a presente tese. Os trabalhos relacionados foram apresentados e discutidos incluindo o destaque dos tópicos de pesquisa encontrados em aberto. Dessa forma, visa-se destacar as principais contribuições e características das soluções propostas nesta tese, que serão detalhadas no decorrer da mesma. Especificamente, no próximo capítulo, serão apresentadas as abordagens propostas para análise e avaliação de grupos de pesquisadores.

3 ANÁLISE E AVALIAÇÃO DE GRUPOS DE PESQUISADORES

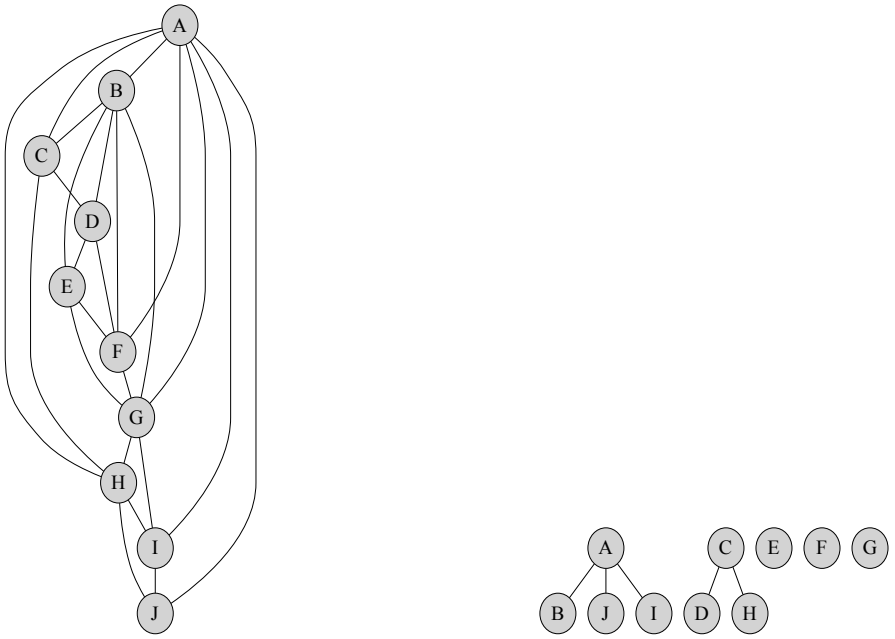
Como mencionado no capítulo anterior, no contexto acadêmico, métricas, principalmente bibliométricas, vêm sendo estudadas e desenvolvidas. Especificamente, métricas para analisar grupos de pesquisadores vêm sendo estudadas com propósitos de avaliação de qualidade. Além disso, métricas de qualidade são empregadas para definição de rankings, tais como: rankings de periódicos e conferências com base na qualidade de seu corpo editorial e comitê de programa, ranqueamento de universidades com base na qualidade de seus pesquisadores e membros do corpo docente, e raqueamento de propostas de projetos de pesquisa com base na qualidade de seus pesquisadores proponentes (MOLINARI; MOLINARI, 2008; REN; TAYLOR, 2007; SILVA et al., 2010; YAN; LEE, 2007). Tendo em vista o elevado número de pesquisas e publicações em bibliometria, resolveu-se explorar outra faceta para definição de indicadores de avaliação de qualidade de grupos de pesquisadores. Nesta tese, resolveu-se aliar a tendência atual de estudos em análises de redes sociais e aplicá-la na definição de tais indicadores. É importante notar que unindo as poderosas análises dadas por SNA e os dados disponibilizados sobre comunidades de pesquisa, podem ser definidas métricas para avaliar a forma como os grupos de pesquisa colaboram. Mais ainda, SNA permite analisar colaborações entre pesquisadores, bem como quantificar comportamentos de interação científica, que, por sua vez, são facetas que podem ser exploradas para propósitos de avaliação de qualidade.

Este capítulo detalha as abordagens propostas nesta tese em relação à análise e avaliação de grupos de pesquisadores no contexto acadêmico. Inicialmente, é apresentado o coeficiente de Gini aplicado à análise de redes sociais (seção 3.1). Posteriormente, são apresentadas as possíveis métricas associadas a indicadores de qualidade, considerando a questão social para ranquear grupos de pesquisadores (seção 3.2). Por fim, são mostrados os experimentos efetuados para validar e avaliar às abordagens propostas (seção 3.3).

3.1 Coeficiente de Gini aplicado à Análise de Redes Sociais

Nesta tese, é proposto que o coeficiente de Gini seja usado como métrica (como aquelas apresentadas na seção 2.1.2) dentro de Análise de Redes Sociais. São propostas duas diferentes formas de análise nomeadas: g_e e g_c. Nessas formas, a diferença consiste na distribuição adotada para análise da rede social, a partir da qual o coeficiente de Gini (descrito previamente na seção 2.2.3) é calculado.

A primeira delas, nomeada g_e, define o coeficiente de Gini calculado considerando que os vínculos relacionais (arestas) entre atores (nós) podem ser vistos como uma distribuição dos possíveis pares de atores dentro de uma rede (cada ator combinado com todos



(a) (b)

Figura 3.1: Exemplos de Redes Sociais: (a) rede conectada, e (b) rede pobremente conectada.

os outros atores da SN). Para cada par de atores, o valor associado é ou 0 (zero) quando não existe relacionamento entre eles, ou o peso normalizado do relacionamento entre eles (1 em uma SN não ponderada). A distribuição g_e é perfeitamente igualitária quando as relações entre todos os pares de atores têm o mesmo peso. Por outro lado, ela será completamente desigual quando apenas um par de atores representar todo o relacionamento da SN. Um baixo valor de coeficiente de Gini indica uma distribuição mais igualitária, e um alto valor indica uma distribuição mais desigual.

A segunda forma, nomeada g_c, define o coeficiente de Gini baseado no grau de conectividade dos atores da rede. Nesse caso, o foco é mensurar a distribuição de relacionamentos entre atores. Dessa forma, a distribuição é formada pelo número de conexões de cada ator na rede social SN. Complementarmente, para g_c, altos valores de coeficiente de Gini indicam que a conectividade da SN é homogênea, enquanto baixos valores revelam que a conectividade da SN é não homogênea (por exemplo, as conexões ocorrem somente entre poucos atores, enquanto que a maioria não está conectada).

Por exemplo, a Figura 3.1(a) e (b) apresenta duas redes sociais distintas para um grupo de atores. Para tais redes (a) e (b), os valores de g_e são, respectivamente, 0,7108 e 0,9865; e os valores de g_c são, respectivamente, 0,1283 e 0,2724. Pode-se observar que, a rede na Figura 3.1(a) está relativamente bem conectada e seus relacionamentos são bem distribuídos. Por outro lado, a rede na Figura 3.1(b) é pobremente conectada, com poucas relações considerando a rede como um todo. É importante notar que esse exemplo enfatiza como os valores de coeficiente de Gini, de fato, caracterizam as distribuições das relações dessas redes.

Function QualityAssessment
Input: A list of research groups G
1. **for all** $g_i \in G$ **do**
2. R_i = Get the list of researchers that are currently members of g_i
3. **for all** $r_j \in R_i$ **do**
4. P_j = Get the list of publications of r_j from digital library
5. **end for**
6. E_i = Find the co-authorship relations within P
7. SN_i = Materialize the social network $SN(R_i, E_i)$
8. Apply SNA metrics to SN_i
9. **end for**
10. **return** $Ranking(SN)$

Figura 3.2: Função para avaliação de qualidade de grupos de pesquisa.

3.2 Métricas de qualidade para ranquear grupos de pesquisadores

Nesta seção, são propostas uma nova função para avaliação de qualidade de grupos de pesquisadores (seção 3.2.1) e novas métricas e análises para avaliação de qualidade em redes sociais (seção 3.2.2).

3.2.1 Função Geral

A Figura 3.2 descreve uma função geral para avaliação de qualidade de grupos de pesquisadores. Dada uma lista de grupos de pesquisa, lembrando que se quer ranquear eles ao final, obtém-se a lista de seus pesquisadores (linhas 1-2), e o conjunto de publicações de cada um desses pesquisadores (linhas 3-5). Uma vez que se tem os pesquisadores e suas publicações, o próximo passo é casar seus nomes e suas publicações, verificando as relações de coautoria (linha 6). A lista de pesquisadores e suas coautorias define a rede social (linha 7), que por sua vez é analisada com métricas de análise de redes sociais (linha 8). Finalmente, um ranking dos grupos é produzido como saída da função (linha 10).

3.2.2 Novas Métricas

As novas métricas são baseadas em características desejadas em grupos de pesquisa de alta qualidade. A primeira é a Eficiência Social (seção 3.2.2.1) que se baseia na necessidade de um comportamento colaborativo no grupo e a necessidade de não-existência de indivíduos "socialmente ineficientes". A segunda é a análise do maior autovalor (seção 3.2.2.2) proposto para inferir qualidade com base na necessidade de um número grande de bons pesquisadores e de alta densidade de colaborações. A terceira é o índice β (seção 3.2.2.3) proposto para inferir qualidade, com base na necessidade de uma alta média de artigos em coautoria e, simultaneamente, que seus pesquisadores tenham um comportamento similar em relação à conectividade. As métricas são definidas como segue.

3.2.2.1 Eficiência Social

A eficiência social (*social efficiency*) objetiva medir o percentual de nós que contribuem para as conexões da rede. A eficiência social, nomeada como se, é apresentada na Equação 3.1, definida como o valor 1 subtraído do valor da ineficiência social. A ineficiência social (*social inefficiency*) é calculada pelo número de nós sem arestas na rede social sendo analisada, dividido pelo número total de nós da rede, como apresentado na Equação 3.2. Para definição desta, a Equação auxiliar 3.3 é utilizada, onde dc refere-se ao grau de centralidade (conforme Equação 2.1 previamente apresentada).

$$se(G) = 1 - si(G) \tag{3.1}$$

onde:

$$si(G) = \frac{1}{N} \sum_{i=1}^{N} E(n_i) \tag{3.2}$$

$$E(n_i) = \begin{cases} 1, & \text{se } (dc(n_i) = 0) \\ 0, & \text{caso contrário} \end{cases} \tag{3.3}$$

3.2.2.2 Maior autovalor

Os autovetores de uma matriz quadrada são vetores não-nulos que, depois de serem multiplicados pela matriz, permanecem paralelos ao vetor original[1]. Para cada autovetor, o correspondente autovalor é o fator pelo qual o autovetor é escalado quando multiplicado pela matriz. Dada a expressão matemática: $Av = \lambda v$, tal que se A é uma matriz quadrada, o vetor não-zero v é um autovetor de A se existe um escalar λ. O escalar λ é dito ser o autovalor de A correspondente ao v.

Propriedades espectrais de grafos têm apontado formas interessantes para descrever suas topologias (NAJIMINAINI; SUBEDI; TRAJKOVIC, 2009). Entretanto, estas propriedades podem também ser usadas para medir o nível de interação das pessoas representadas em um grafo. O maior autovalor (*highest eigenvalue*) da matriz de adjacência de um grafo (ou até mesmo de uma matriz de adjacência valorada, onde as arestas descrevem o nível de interação entre pares de autores) tem duas características que ajudam a classificar o nível de interação do grupo definido por sua topologia: matrizes de maiores dimensões e grande densidade de arestas podem levar a ampliar o maior autovalor.

A hipótese é que essas duas características podem definir grupos de pesquisadores de alta qualidade: (i) grande número de "bons" pesquisadores; e (ii) alta densidade de arestas, que significa um alto nível de publicações em conjunto, ou até mesmo uma boa comunicação entre pesquisadores do mesmo grupo, e não somente desses pesquisadores com membros de grupos externos.

Com base em tal hipótese, são propostas duas métricas interessantes para quantificar a qualidade de grupos de pesquisadores que são o maior autovalor da matriz de adjacência binária (*binary adjacency matrix*) e da matriz de adjacência valorada (*valued-adjacency matrix*).

3.2.2.3 Índice β

Para propósitos de ranqueamento, a hipótese considerada para determinação da métrica, apresentada nesta seção, é de que as duas características a seguir devem ocorrer, simultaneamente, em grupos de pesquisa de qualidade: (i) alta média de artigos em coautoria e (ii) que seus autores tenham um comportamento similar em relação à conectividade. Esse comportamento é refletido por valores altos do índice β, que será apresentado a seguir.

Referente à primeira característica desejada, a métrica nomeada ρ é proposta para calcular a média do número de artigos em coautoria entre pares de pesquisadores cooperantes, conforme definido na Equação 3.4.

[1] Eigenvalues and Eigenvectors - Wikipedia: http://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors

$$\rho = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} n_{ij}}{e} \tag{3.4}$$

onde n_{ij} denota o número de artigos em conjunto entre o par de vizinhos $< i, j >$, n denota o número total de autores, e e denota o número total de arestas existentes na Rede Social SN (ou seja, SN considera somente os pares de pesquisadores $< i, j >$ que têm pelo menos um artigo em coautoria).

Referente à segunda característica desejada, a métrica do coeficiente de Gini g_c (apresentada previamente na seção 3.1) é aplicada. Essa foi escolhida porque é mais adequada para análises comparativas entre diferentes redes sociais de pesquisa, nas quais as redes provavelmente tenham tamanhos diferentes, porque são compostas por diferentes conjuntos de pesquisadores.

Tendo-se as duas métricas anteriores definidas, a seguir é apresentado o índice β que é proposto para combinar ambas, a fim de avaliar redes de colaboração de forma justa para propósitos de ranqueamento, conforme Equação 3.5.

$$\beta = \frac{\rho}{g_c} \tag{3.5}$$

onde ρ é a média de artigos em coautoria da rede social SN e g_c é o coeficiente de Gini, considerando a distribuição da conectividade desta mesma rede SN, ou seja, analisando a distribuição do número de coautores de cada pesquisador na SN.

3.2.2.4 Exemplo de medição das métricas

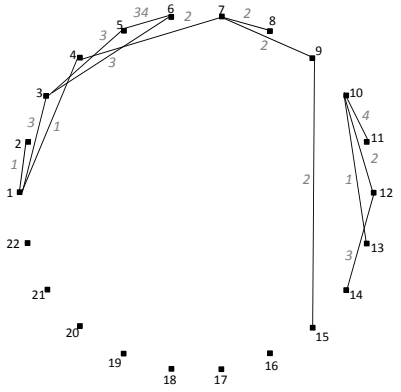
Como exemplo, é apresentado o cálculo das métricas para uma rede social simples e genérica. A Figura 3.3(a) ilustra uma rede simples e genérica para representar uma rede de pesquisa composta por 22 pesquisadores, que são representados pelos pontos numerados na figura. Nessa figura, as coautorias entre pesquisadores são representadas por linhas contínuas ligando pontos e os pesos das arestas são representados pelos números em cinza na figura, indicando o número de artigos em coautoria entre os pesquisadores. A Figura 3.3(b) apresenta uma tabela com os respectivos valores encontrados pelo cálculo das diferentes métricas com base nessa rede. As métricas apresentadas incluem eficiência social, maior autovalor (matriz de adjacência binária e ponderada pelo número de artigos em comum), média de publicações em coautoria entre pares cooperantes (ρ), coeficiente de Gini (g_c) e índice β (discutidas nas seções 3.1 e 3.2).

Embora seja difícil dizer se este grupo de pesquisa é top ou não, avaliando apenas os valores individuais apresentados na Figura 3.3, esses valores servem aqui apenas como exemplo da aplicação das métricas. Na próxima seção, será discutido como esses valores podem ser aplicados para avaliação de qualidade e com propósitos de geração de *ranking*.

3.2.3 Aplicação de métricas para avaliação de qualidade e para propósitos de geração de *ranking*

Como base na hipótese de que um bom comportamento é a alta interação cooperativa entre pesquisadores, as redes de coautoria de grupos de pesquisa *top* devem ser tão conectadas quanto possível. Para avaliar tal comportamento e tentar inferir a qualidade, são usadas as métricas apresentadas nas seções 2.1.2 e 3.2.

Além de avaliar os grupos de pesquisa, as métricas são usadas para construir rankings ordenados por qualidade. Para todas as métricas, quanto maior seu valor, maior a quali-



(a)

Métricas	Valores
Eficiência Social	0,68
Maior autovalor (matriz adj. binária)	2,32
Maior autovalor (matriz adj. valorada)	34,52
Média de publicações em coautoria (ρ)	4,5
Coeficiente de Gini (g_c)	0,5524
Índice β	0,1943

(b)

Figura 3.3: Exemplos de (a) rede social e (b) respectivos valores das métricas.

dade da rede analisada. Esse é o caso das seguintes métricas: densidade, coeficiente de clusterização, coeficiente gigante, eficiência social, maior autovalor, índice β e média ρ. Estas métricas devem ser ordenadas decrescentemente a fim de definir o ranking. A única exceção é a métrica do coeficiente de Gini g_c, que deve ser ordenada crescentemente. Maiores detalhes sobre os resultados da sua aplicação para fins de ranking, com base em qualidade, são discutidos na seção 3.3.2. Além disso, um exemplo da aplicação destas e outras métricas para fins de ranqueamento são apresentadas na seção 3.3.2.3.

3.3 Avaliação Experimental

Esta seção apresenta a avaliação experimental efetuada. Especificamente, são apresentados os experimentos relativos à análise da Evolução Temporal de uma Rede Social com o uso do coeficiente de Gini (seção 3.3.1) e à avaliação de qualidade propriamente dita (seção 3.3.2).

3.3.1 Evolução Temporal de uma Rede Social analisada com o coeficiente de Gini

Como apresentado na seção 3.1, são propostas diferentes formas de aplicar o coeficiente de Gini na análise de redes sociais. Nesta seção, tais propostas são aplicadas em uma rede social de coautoria, na qual atores são pesquisadores e vínculos relacionais são colaborações de pesquisa, expressas através de artigos em coautoria, entre pares de pesquisadores. Além disso, é considerada uma rede social representada por um grafo ponderado no qual os pesos são dados pela razão entre o número de artigos publicados em coautoria normalizado pelo total de publicações de um dos autores, de acordo com a direção da relação (para detalhes, ver Equação 4.9 apresentada no próximo capítulo). A seguir, é apresentado o conjunto de dados usado para construir as redes sociais neste experimento e as respectivas análises do coeficiente de Gini.

3.3.1.1 Descrição do conjunto de dados

Neste experimento, é utilizada uma rede composta por 27 pesquisadores envolvidos no InWeb, Instituto Nacional de Ciência e Tecnologia para Web[2]. Esse projeto começou em 2008 e todos os seus pesquisadores são membros do corpo docente, incluindo professores titulares, associados, adjuntos e assistentes, nas instituições de ensino e pes-

[2]InWeb: http://inweb.org.br

Tabela 3.1: As instituições participantes do projeto InWeb.

#Id	Instituições
1	UFAM, Manaus, AM, Brasil
2	UFRGS, Porto Alegre, RS, Brasil
3	CEFET-MG, Belo Horizonte, MG, Brasil
4	UFMG, Belo Horizonte, MG, Brasil

quisa brasileiras participantes (UFMG, UFRGS, UFAM e CEFET-MG) com programas de Pós-graduação em Ciência da Computação.

O conjunto de dados, utilizado para construir as redes, foi obtido da DBLP em 03 de agosto de 2010. A fim de avaliar a evolução das relações sociais, foram consideradas duas redes construídas utilizando diferentes intervalos de tempo. O primeiro intervalo de tempo inclui as publicações dos pesquisadores até 2007, definindo uma rede referenciada como SN2007. O segundo intervalo de tempo inclui publicações até 2010, definindo uma rede chamada SN2010. O projeto começou em 2008, então SN2007 reflete as colaborações entre os pesquisadores antes do início do projeto InWeb, enquanto que SN2010 mostra as colaborações dos pesquisadores até dois anos depois do início (durante o desenvolvimento do projeto). A seguir é discutido como o coeficiente de Gini pode ser usado para avaliar o impacto de ter-se o projeto, nas colaborações entre seus pesquisadores membros.

A Figura 3.4 mostra a evolução ocorrida nas colaborações entre os pesquisadores considerando um comparativo entre SN2007 e SN2010. Ao invés de apresentar uma figura para cada rede, ambas foram reunidas na Figura 3.4, conforme segue. Os números dentro dos nós especificam os pesquisadores: o primeiro número (#Id) identifica a instituição de afiliação dos pesquisadores de acordo com a Tabela 3.1, e o segundo número identifica cada pesquisador dentro da instituição. Os pesquisadores que têm pelo menos um artigo em coautoria estão conectados pelas arestas: linhas contínuas em cinza representam conexões que não foram intensificadas durante o projeto InWeb; linhas contínuas em preto representam as conexões intensificadas durante o desenvolvimento do projeto (de 2008 até 2010); e as linhas tracejadas representam as novas conexões entre pesquisadores iniciadas durante o projeto até o ano de 2010.

3.3.1.2 Avaliações

Esta seção avalia as redes SN2007 e SN2010 sob diferentes perspectivas: análise global, análise cooperativa e curvas de Lorenz.

Análise Global. A primeira análise considera a distribuição de pesos de todos os possíveis vínculos relacionais entre pesquisadores, incluindo o peso zero entre pesquisadores que não têm nenhum artigo em coautoria. Essa análise do coeficiente de Gini entre todos os pares de pesquisadores é importante para avaliar a colaboração global entre os pesquisadores da rede social estudada (proposta de uso do coeficiente de Gini nomeada g_e).

Os resultados dessa primeira análise (ver Análise 1 da Tabela 3.2) mostram que o valor do coeficiente de Gini é menor em SN2010 (SN2010_allPairs) do que em SN2007 (SN2007_allPairs). Esses resultados são coerentes e eles são justificados porque o SN2010 é mais conectado que o SN2007. Entretanto, a diferença entre os valores do coeficiente de Gini de ambos é muito pequena e insignificante estatisticamente. É possí-

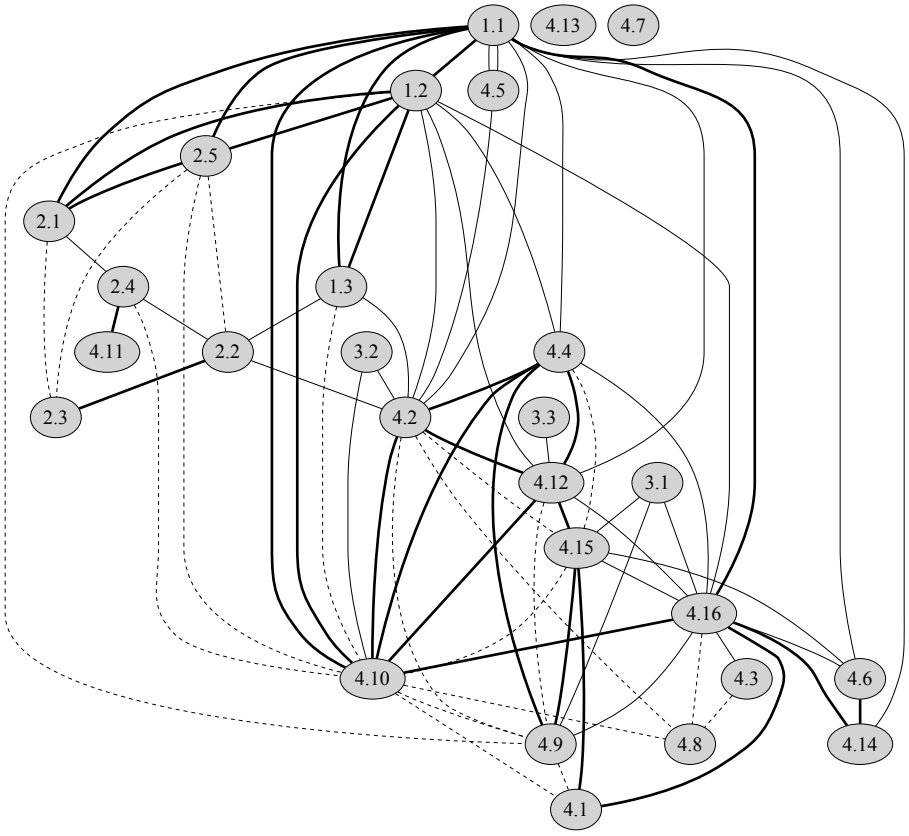


Figura 3.4: Comparativo entre a Rede Social do InWeb antes do início do projeto e durante seu desenvolvimento (com base em dados da DBLP): linhas cinzas para conexões não intensificadas, linhas pretas para conexões intensificadas, e linhas tracejadas para novas conexões.

vel notar que valores altos de coeficiente de Gini foram obtidos em ambas as redes. Esses resultados indicam distribuições desiguais, significando que essa rede de colaboração do InWeb está ainda muito desconectada, em relação a uma rede totalmente conectada, em ambos os intervalos de tempo considerados.

O projeto de pesquisa tem duração de cinco anos e ainda está em desenvolvimento. No final dos cinco anos, o objetivo é que se tenha aumentado o nível de cooperação entre os pesquisadores envolvidos. Do ponto de vista social, esse tipo de análise é fundamental para entender as características e propriedades da SN sob estudo.

Análise Cooperativa. A segunda análise considera apenas os pares de pesquisadores que tiveram artigos em coautoria (pesos não nulos) em uma das SNs sendo analisadas (SN2007 ou SN2010). Essa análise do coeficiente de Gini considerando somente os pares cooperativos de pesquisadores é importante para avaliar o nível genuíno de colaboração entre estes pares que têm vínculos relacionais modelados pela SN.

Pares cooperativos até 2007. Os resultados da análise, considerando apenas os pares cooperativos de pesquisadores até 2007 (ver Análise 2.1 da Tabela 3.2), mostram que o valor do coeficiente de Gini é menor em SN2010 (SN2010_Pairs2007) do que em SN2007 (SN2007_Pairs2007). Essa diferença reflete a melhoria na homogeneidade da distribuição de pesos entre pesquisadores que já tinham colaborado antes do início do projeto InWeb.

Tabela 3.2: Valores do coeficiente de Gini analisados nas redes de colaboração do InWeb.

Análise	Distribuição	Coeficiente de Gini
(1)	SN2007_allPairs	0,9471
	SN2010_allPairs	0,9327
(2.1)	SN2007_Pairs2007	0,5824
	SN2010_Pairs2007	0,5735
(2.2)	SN2007_Pairs2010	0,7009
	SN2010_Pairs2010	0,6160

As colaborações entre os pesquisadores que foram intensificadas, como pode ser observado na Figura 3.4, contribuíram para uma maior igualdade na distribuição em SN2010 do que em SN2007.

Pares cooperativos até 2010. Os resultados da análise, considerando apenas os pares cooperativos de pesquisadores até 2010 (ver Análise 2.2 da Tabela 3.2), mostram que o valor do coeficiente de Gini é menor em SN2010 (SN2010_Pairs2010) do que em SN2007 (SN2007_Pairs2010). Mais uma vez, essa diferença reflete as melhorias ocorridas nas colaborações dos pesquisadores depois do início do projeto InWeb. Isso ocorre porque os pares de autores considerados em ambas as SNs são iguais. É possível notar que as colaborações entre os pesquisadores foram intensificadas, como pode ser observado na Figura 3.4, e o valor do coeficiente de Gini identificou essa intensificação da conectividade. Mais ainda, é possível perceber que altos valores foram obtidos para ambas as redes. Esses resultados indicam distribuições desiguais em ambas as SNs. Em resumo, é possível afirmar que: (i) em SN2007, existem pares de pesquisadores que não estão conectados e a distribuição de pesos não é igualitária, e (ii) em SN2010, algumas colaborações existentes são intensificadas (aumentaram os pesos) enquanto que novas colaborações emergiram (primeiras publicações em coautoria, provavelmente com baixos pesos), fazendo a distribuição dos pesos ainda não igualitária em valores.

Curvas de Lorenz. Como apresentado anteriormente, o coeficiente de Gini é calculado como o percentual da área entre a linha de perfeita igualdade (inclinação de 45^o) e a curva de Lorenz observada em relação à área entre a linha de perfeita igualdade e a linha de perfeita desigualdade. As representações gráficas das curvas de Lorenz das distribuições analisadas neste estudo de caso são apresentadas na Figura 3.5. Nessa figura, o percentual cumulativo dos pares de pesquisadores é plotado no eixo x; o percentual cumulativo dos pesos dos relacionamentos, no eixo dos y. A legenda das distribuições é a mesma usada na Tabela 3.2.

Na segunda análise (cooperativa), somente os pares de pesquisadores que têm algum vínculo relacional entre eles foram selecionados. Nesse caso, é possível notar que existem pequenas áreas entre as curvas de Lorenz das distribuições e a linha de perfeita igualdade, diferente dos resultados com as áreas correspondendo a primeira análise (global). Em outras palavras, esses valores indicam distribuições mais igualitárias do que na primeira análise, isto é, SNs mais conectadas. Tais resultados são coerentes pelos conjuntos de dados sendo analisados.

Como um resumo do primeiro conjunto de experimentos, o coeficiente de Gini pode ser empregado para analisar uma rede social de pesquisa real. Os resultados mostram evidências significativas da validade e aplicabilidade do coeficiente de Gini no contexto

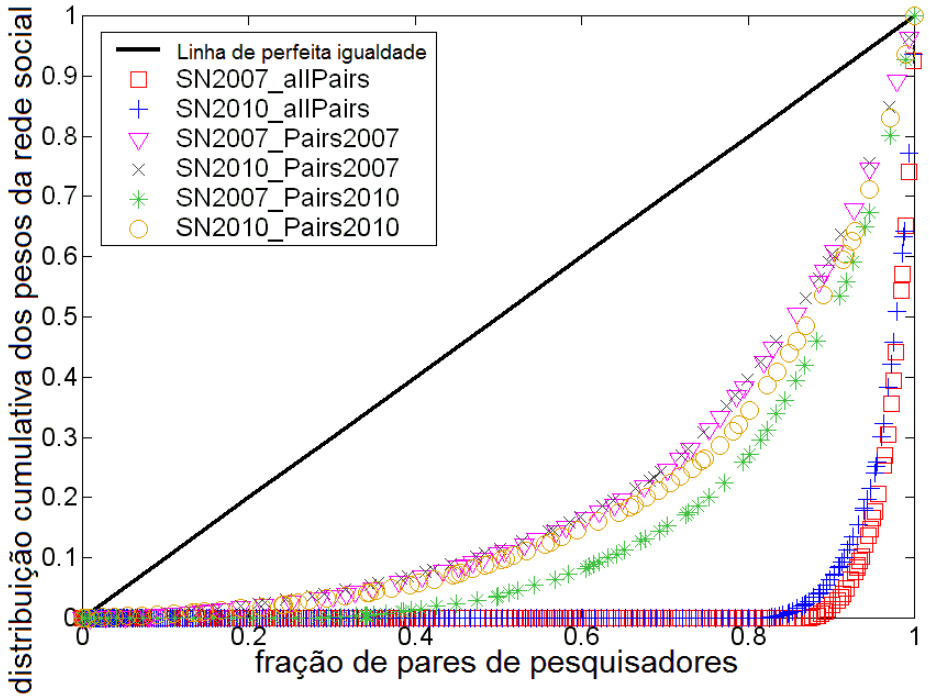


Figura 3.5: Curvas de Lorenz para as distribuições das redes de colaboração do InWeb.

de Redes Sociais. Essa abordagem é um método promissor para a avaliação *a posteriori* de projetos científicos de cooperação e mobilidade do ponto de vista da análise da cooperação científica.

3.3.2 Experimentos sobre avaliação de qualidade

Para avaliar a qualidade de grupos de pesquisa, um grande desafio é como estabelecer um *baseline* com o qual o resultado obtido (pelo novo ranking) poderá ser comparado. No Brasil, a cada triênio, a CAPES avalia os programas de Pós-graduação do país e atribui uma nota de 1 a 7. Tal nota estabelece um ranking de todos os programas. Aproveitando a existência de tal ranking oficial, nesta seção as métricas existentes bem como as novas métricas são avaliadas considerando os grupos de pesquisa representados por cada programa de Pós-graduação. Após, o ranking obtido com tais métricas pode ser avaliado considerando o *baseline* do ranking oficial.

Com estes experimentos, quer-se mostrar que as novas métricas (eficiência social, maior autovalor e índice β) são apropriadas para avaliação de qualidade de grupos de pesquisa (nesse caso, programas de Pós-graduação). Também foi discutido (nos trabalhos relacionados da seção 2.2.2) que uma única faceta, muitas vezes, não é suficiente para avaliação de qualidade. Portanto, é bastante razoável que não se faça uso de nenhuma delas isoladamente como *baseline* para comparativo dos resultados. Por isso, nesta seção, apresentamos a avaliação da CAPES[3] como *baseline*.

3.3.2.1 Ranking da CAPES como baseline

A CAPES é uma agência federal brasileira para qualificação de recursos humanos e é responsável pela avaliação oficial dos programas de Pós-graduação das universidades. A avaliação é desenvolvida considerando um período de três anos. Para cada três anos de

[3]CAPES: http://www.capes.gov.br/

avaliação, os programas de Pós-graduação são classificados em uma escala *Likert* de sete níveis (de 1 a 7). A classificação é feita por especialistas e é desenvolvida com base em uma série de critérios de qualidade tais como: proposta do programa, membros do corpo docente, estudantes e teses desenvolvidas, produção intelectual (publicações), e inserção social. A agência tem um complexo conjunto de equações a fim de ponderar todas essas facetas. Os níveis de qualidade mais altos são 7 e 6 e representam programas de Pós-graduação com desempenhos comparáveis aos melhores programas internacionais. Para garantir a transparência do processo, todos os dados são divulgados e disponibilizados na página Web da CAPES[4]. Foi selecionado um conjunto de dados composto por 27 programas de Pós-graduação em Ciência da Computação brasileiros. Esse conjunto de programas inclui todos os programas de níveis 7, 6 e 5 (de acordo com o ranking mais atual publicado para o triênio 2007-2009), e um conjunto selecionado de programas dos níveis 4 e 3 (abaixo do nível 3, os programas necessitam promover uma reestruturação ou serão fechados). Esse conjunto de dados é utilizado como *baseline* e é apresentado na Tabela 3.3, na qual a primeira coluna corresponde ao nome do programa de Pós-graduação selecionado e a segunda coluna apresenta a respectiva última avaliação tri-anual (período 2007-2009) desenvolvida pela CAPES.

3.3.2.2 Descrição do conjunto de dados

O conjunto de dados utilizado nestes experimentos inclui os pesquisadores de 27 programas de Pós-graduação em Ciência da Computação brasileiros (da Tabela 3.3) e suas publicações. Embora tais programas sejam considerados, esta avaliação pode ser aplicada em programas de quaisquer outras áreas, tais como física e medicina.

No total, foram considerados 732 pesquisadores, isto é, membros do corpo docente incluídos na última avaliação da CAPES. Seus dados de publicações foram extraídos da DBLP[5] em 03 de agosto de 2010. Foram considerados somente os artigos publicados em anais de conferências ou em periódicos indexados pela DBLP até 2009, uma vez que a avaliação da CAPES considera apenas publicações de 2007 a 2009. A justificativa é que para avaliar a qualidade de um grupo de pesquisa ou de um programa de Pós-graduação, que é o caso desta avaliação experimental, apesar de considerar um período pré-determinado, poderia ser ingênuo demais considerar somente dados daquele período restrito para estimativa de qualidade. Inclusive, para a avaliação da CAPES, um programa pode subir ou descer de nível de uma avaliação trienal para outra, mas o histórico passado ("reputação" do programa construída ao longo dos anos) também é um fator que acaba sendo levado em consideração. Além disso, pelo uso dos dados da DBLP, em que existe certa demora para certas publicações serem indexadas, como a coleta dos dados para os experimentos foi realizada em 2010, algumas publicações do período até 2009, provavelmente, ainda não tivessem sido indexadas. Mais ainda, a DLBP indexa apenas as publicações mais relevantes que foram publicadas em eventos de algum impacto internacional (publicações de brasileiros em eventos como mostras de produção universitária, escolas regionais brasileiras ou eventos nacionais de menor visibilidade internacional não estão indexadas nesta). Dessa forma, objetivando uma maior quantidade de dados para embasar uma avaliação dos programas de forma mais consistente, com base em colaborações, foram consideradas todas as publicações dos pesquisadores indexadas na DBLP até 2009 (que para muitos pesquisadores corresponderá a um subconjunto de suas principais publicações).

[4]Avaliação CAPES: http://www.capes.gov.br/avaliacao/avaliacao-da-pos-graduacao

[5]DBLP: http://www.informatik.uni-trier.de/~ley/db

Tabela 3.3: Conjunto selecionado de Programas de Pós-graduação em Ciência da Computação brasileiros e sua respectiva classificação CAPES (de acordo com a avaliação trianual 2007-2009).

Programa de Pós-graduação	**Classificação da CAPES**
COPPE/UFRJ	7
PUC/RIO	7
UFMG	7
UFPE	6
UFRGS	6
UNICAMP	6
USP/SC	6
UFF	5
USP	5
PUC/PR	4
PUC/RS	4
UFAM	4
UFBA	4
UFC	4
UFCG	4
UFES	4
UFPR	4
UFRJ	4
UFRN	4
UFSC	4
UFSCAR	4
UNB	4
UNISINOS	4
PUC/MG	3
UCPEL	3
UFG	3
UFPA	3

Note que estas publicações são necessárias a fim de especificar as relações de coautoria entre pesquisadores. Estudos recentes (LAENDER et al., 2008; REITZ; HOFFMANN, 2010) discutem que a cobertura das subáreas da Ciência da Computação pela DBLP tem atingido valores de aproximadamente 67%, cobrindo acima disto até 96% de algumas subáreas. No entanto, a DBLP é amplamente aplicada para obter publicações da área da Ciência da Computação, muito embora alguns resultados de exceção podem ser motivados pela limitada cobertura para alguma subárea em específico.

3.3.2.3 Avaliação e Resultados

Nestes experimentos, foi seguida a função da Figura 3.2 para avaliação de qualidade de grupos de pesquisa considerando os programas de Pós-graduação supracitados (Tabela 3.3). As redes sociais de coautoria foram materializadas e algumas delas, uma de cada nível de classificação da CAPES, são ilustradas na Figura 3.6. Nessa representação visual, pesquisadores são representados por pontos numerados e os pares desses, que possuem pelo menos um artigo publicado em coautoria, são ligados por linhas. Uma rápida visualização da figura é suficiente para perceber que os programas dos níveis mais altos da classificação CAPES têm mais pesquisadores conectados, isto é, maior interação colaborativa, enquanto que os outros têm um comportamento social/colaborativo mais baixo. Dessa forma, as diferenças entre os programas de níveis 7 e 6 para os outros são normalmente mais aparentes.

Também de forma ilustrativa, os valores avaliados das diferentes métricas, para fins de geração de *ranking*, são apresentados para as Redes Sociais ilustradas na Figura 3.6. Tais valores avaliados para as cinco redes, uma de cada nível de classificação da CAPES, são apresentados na Tabela 3.4. Nessa, cada linha apresenta os resultados de ranqueamento, considerando a métrica correspondente calculada para cada um dos cinco programas exemplificados. Os resultados apresentados incluem o valor da métrica calculada e, entre colchetes, a posição relativa do programa correspondente dentre os cinco considerados, a partir do uso dessa métrica, para fins de ranqueamento. Nesse caso, o *ranking* ideal seria a ordenação dos programas Nível 7, Nível 6, Nível 5, Nível 4 e Nível 3, respectivamente, nas posições [1], [2], [3], [4] e [5]. Esses resultados são apresentados apenas de forma ilustrativa, a avaliação completa (considerando os 27 grupos de pesquisa representados pelos programas de Pós-graduação) será detalhada e os resultados avaliados a seguir.

A fim de avaliar os resultados (considerando os 27 programas de Pós-graduação) em relação ao *baseline*, foi necessária alguma métrica que permitisse comparação entre rankings. Um modo comum de efetuar tal comparação é empregar o coeficiente de Spearman para avaliar a correlação entre rankings. Entretanto, uma vez que a CAPES ranqueia os programas por níveis, todos os programas classificados em um mesmo nível podem ser considerados como "empatados". Por isso, foi utilizada uma variação do coeficiente de Spearman que lida com empates, como apresentado em (SIEGEL; CASTELLAN, 1988). Em resumo: quanto maior o valor do coeficiente de Spearman, maior a correlação entre os rankings sendo comparados. Além disso, o nível de significância dos resultados obtidos é também avaliado. O limiar de significância estatística de 0,01 foi utilizado e os resultados estão apresentados na Tabela 3.5 - que apresenta os resultados ordenados decrescentemente de acordo com o coeficiente de Spearman.

Os resultados do coeficiente de Spearman obtidos pelas métricas tradicionais de SNA (densidade, coeficiente de clusterização global e coeficiente de clusterização ponderado) não foram satisfatórios. O pior resultado do coeficiente de Spearman foi obtido pela mé-

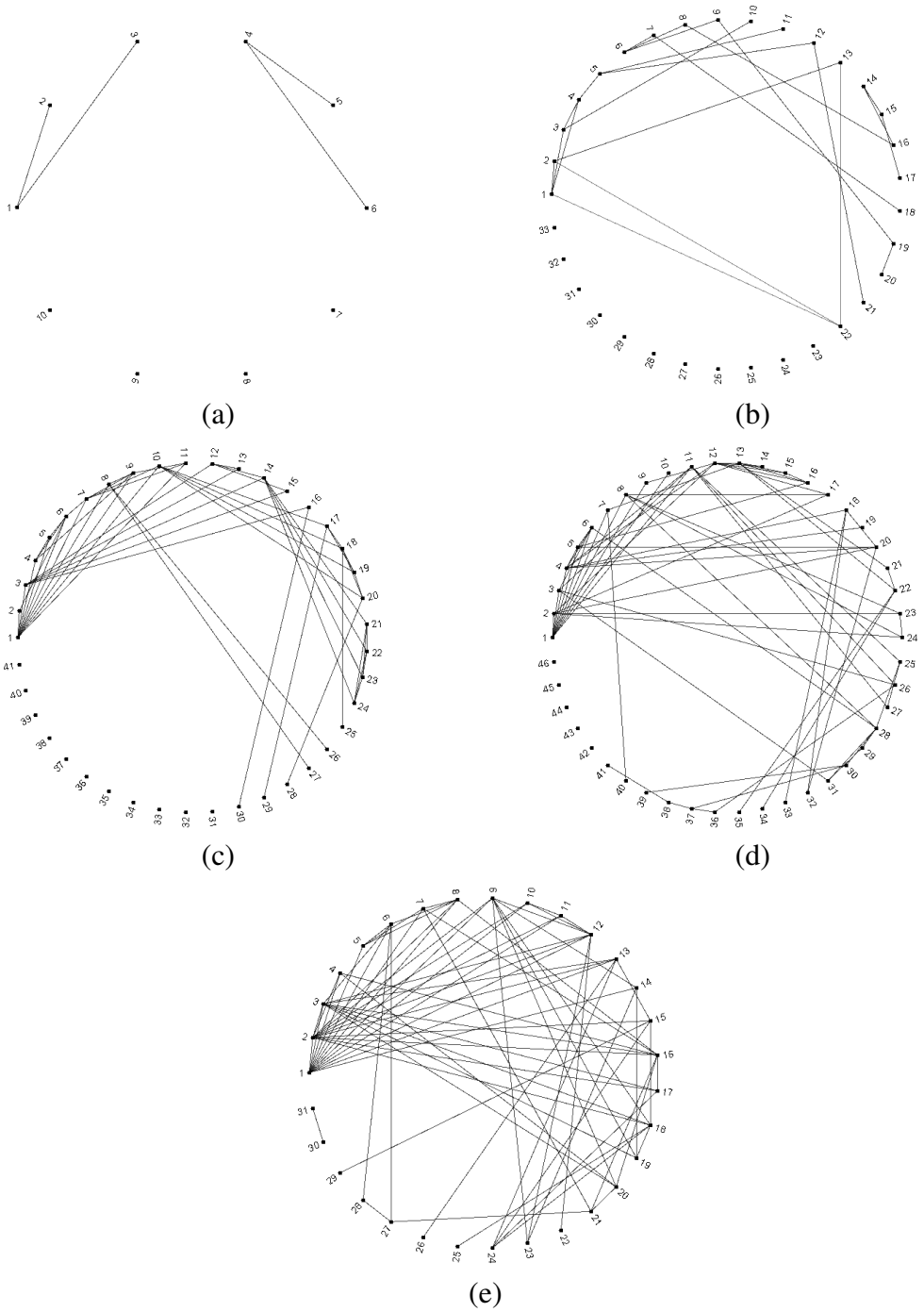


Figura 3.6: Exemplos de Redes Sociais modelando as colaborações internas entre pesquisadores de programas de Pós-graduação. Os programas são classificados pela CAPES como de: (a) Nível 3, (b) Nível 4, (c) Nível 5, (d) Nível 6, e (e) Nível 7.

Tabela 3.4: Exemplo do cálculo das métricas para fins de ranqueamento dos programas de Pós-Graduação exemplificados na Figura 3.6. Cada linha apresenta os resultados de ranqueamento considerando a métrica correspondente calculada para cada um dos cinco programas considerados. Os resultados apresentados incluem o valor da métrica calculada e, entre colchetes, a posição relativa do programa dentre os cinco considerados, a partir do uso dessa métrica para fins de ranqueamento.

Métrica	**Nível 7**	**Nível 6**	**Nível 5**	**Nível 4**	**Nível 3**
Densidade	0,161 [1]	0,069 [3]	0,066 [4]	0,041 [5]	0,089 [2]
Coeficiente de clusterização global	0,484 [2]	0,376 [3]	0,645 [1]	0,151 [4]	0,000 [5]
Coeficiente de clusterização ponderado	0,360 [2]	0,242 [3]	0,495 [1]	0,212 [4]	0,000 [5]
Coeficiente gigante	0,935 [1]	0,891 [2]	0,732 [3]	0,333 [4]	0,300 [5]
Coeficiente gigante - Ineficiência Social	0,935 [1]	0,783 [2]	0,463 [3]	0,000 [4]	-0,100 [5]
Eficiência Social	1,000 [1]	0,891 [2]	0,732 [3]	0,667 [4]	0,600 [5]
Maior autovalor (matriz adj. binária)	7,015 [1]	5,131 [2]	4,891 [3]	2,500 [4]	1,414 [5]
Maior autovalor (matriz adj. valorada)	53,798 [1]	23,309 [2]	22,262 [3]	6,431 [5]	7,071 [4]
Média de publicações em coautoria (ρ)	4,987 [1]	3,211 [2]	2,741 [4]	2,435 [5]	3,000 [3]
Coeficiente de Gini (g_c)	0,399 [5]	0,341 [4]	0,292 [3]	0,227 [2]	0,179 [1]
Índice β	0,054 [1]	0,067 [2]	0,102 [3]	0,157 [5]	0,156 [4]

Tabela 3.5: Resultados do coeficiente de Spearman entre o ranking gerado pela CAPES e os rankings gerados usando diferentes métricas para estimar indicadores de qualidade.

#	**Métricas**	**Coeficiente de Spearman**	**Significância**
1	Maior autovalor (matriz adj. binária)	0,807	Correlacionado
2	Coeficiente gigante - Ineficiência Social	0,736	Correlacionado
3	Maior autovalor (matriz adj. valorada)	0,732	Correlacionado
4	Coeficiente gigante	0,707	Correlacionado
5	Eficiência Social	0,682	Correlacionado
6	Índice β	0,642	Correlacionado
7	Coeficiente de Gini (g_c)	0,422	Não correlacionado
8	Coeficiente de clusterização ponderado	0,386	Não correlacionado
9	Média de publicações em coautoria (ρ)	0,345	Não correlacionado
10	Coeficiente de clusterização global	0,325	Não correlacionado
11	Densidade	0,248	Não correlacionado

trica de densidade. Estas três métricas geram rankings que **não** são correlacionados com a classificação da CAPES (testados para um limiar de significância de 0,01). Isso provavelmente aconteceu porque considerar a rede ideal como uma rede totalmente conectada pode ser muito ambicioso (e irreal) no contexto acadêmico. Na verdade, isso pode depreciar programas de Pós-graduação com um grande número de pesquisadores, para os quais até mesmo com um grande número de conexões pode-se obter um baixo valor de densidade (porque o número de possíveis conexões também é muito alto). Por outro lado, programas de Pós-graduação com um pequeno número de pesquisadores podem obter uma injusta vantagem, porque o número de possíveis conexões é muito restrito. Então, estas três métricas **não** se mostraram adequadas para quantificar a qualidade de programas de Pós-graduação para propósitos de geração de ranking.

Como a Tabela 3.5 mostra, utilizar as métricas média de publicações em coautoria (ρ) e coeficiente de Gini (g_c) sozinhas também não proveram bons resultados. É importante notar que ρ mede a intensidade média dos relacionamentos existentes. Os resultados

mostram que somente esta métrica (ρ) não é suficiente para ordenar corretamente os programas de Pós-graduação, considerando a classificação da CAPES como *baseline*. Além disso, utilizar o coeficiente de Gini (g_c) sozinho também não é suficiente para esse propósito. Os resultados do coeficiente de Spearman mostram que os rankings gerados por ambas **não** são correlacionados como a classificação da CAPES.

Entretanto, como os resultados do coeficiente de Spearman atestam, quando as duas métricas (ρ e g_c) são combinadas no índice β, o ranking gerado para os programas de Pós-graduação é correlacionado com a avaliação da CAPES (testado a um limiar de significância de 0,01). Esses resultados mostram evidências de que os melhores programas de Pós-graduação são aqueles nos quais seus pesquisadores têm um comportamento colaborativo com ambos: alta intensidade (alto ρ) e distribuições mais homogêneas (baixo g_c), ou seja, alto valor de índice β. Esses resultados corroboram a hipótese de que o comportamento esperado em grupos de pesquisa de alta qualidade seja de que haja uma alta média de artigos em coautoria com a maioria dos pesquisadores tendo o mesmo comportamento em relação à conectividade (baixo coeficiente de Gini).

A métrica de maior autovalor foi calculada para duas matrizes de adjacência representando os pesos das redes sociais sendo analisadas: (i) matriz de adjacência binária, cujos pesos são binários (1 para presença e 0 para ausência de artigos em coautoria entre pares de pesquisadores) e (ii) matriz de adjacência valorada, cujos pesos representam o número de artigos em coautoria entre pares de pesquisadores.

As métricas de coeficiente gigante, eficiência social, diferença entre coeficiente gigante e ineficiência social, e maior autovalor (para ambas as matrizes de adjacência, binária e valorada) apresentaram resultados adequados para ranquear os grupos de pesquisa representados pelos programas de Pós-graduação. As correlações com a classificação da CAPES foram verificadas para o limiar de significância de 0,01. Todas essas métricas obtiveram valores adequados de coeficiente de Spearman. Especialmente, o melhor resultado (aproximadamente 0,807) foi obtido pela métrica de maior autovalor utilizando a matriz de adjacência binária. Para definição de métricas automatizadas de avaliação de qualidade, a principal característica desejada é a simplicidade e objetividade. Então, dentre as métricas apresentadas, a eficiência social, o maior autovalor e o índice β são mais adequadas uma vez que elas são simples de calcular e obtêm resultados satisfatórios.

Uma análise complementar foi realizada para investigar de perto os melhores resultados do maior autovalor usando uma matriz de adjacência binária. Essa análise foi realizada a fim de determinar uma função para representar a variação do maior autovalor pelo nível CAPES (apresentada na Figura 3.7). Especificamente, foram separados os resultados do maior autovalor por nível CAPES (de 3 a 7). Foi calculada a média e o erro padrão dos valores para cada nível, como ilustrado pelos resultados na Figura 3.7. Note que a média é representada por um círculo e o intervalo do erro padrão pelas linhas ao redor do círculo. Então, foi plotado o maior autovalor (*highest*) pelo nível da classificação CAPES (*level*) e encontrou-se uma linha ajustada determinada por $highest = -2.5 + 1.3 * level$. Essa análise mostra que existe um comportamento linear dos maiores autovalores (determinando a qualidade da colaboração interna) e o nível da classificação da CAPES.

Neste capítulo, foi apresentada uma nova abordagem para avaliação de qualidade de grupos de pesquisadores, no contexto acadêmico, utilizando uma faceta social. Especificamente, tal abordagem foi desenvolvida para avaliar grupos de pesquisa com base na qualidade de seu grupo de pesquisadores, inferida com base no comportamento de colaboração entre eles. Tais colaborações foram avaliadas através de diferentes métricas propostas para análises em redes sociais. A abordagem desenvolvida pode ser facilmente

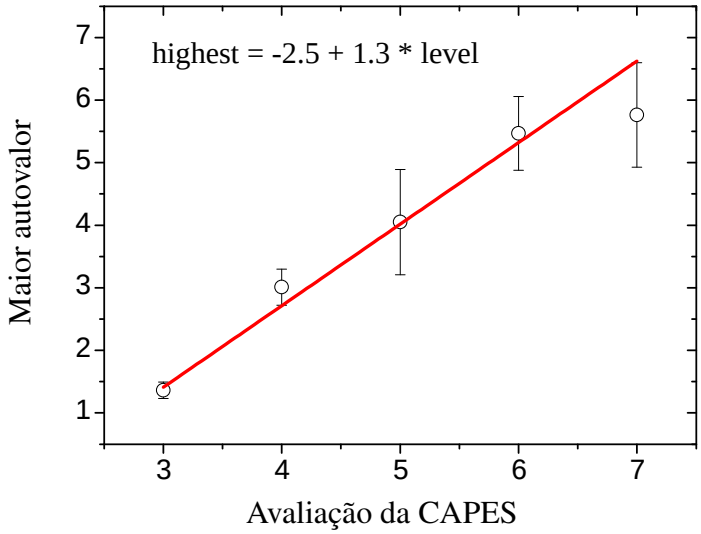


Figura 3.7: Maior autovalor pelo nível de classificação da CAPES.

reproduzida e customizada para avaliação de qualquer grupo de indivíduos no qual o comportamento colaborativo é uma característica desejável e relacionada à qualidade. Além disso, os experimentos apresentados mostraram novas formas de analisar redes sociais e apresentaram as métricas que se mostraram mais adequadas para gerar rankings dos programas de Pós-graduação (correlacionados com o ranking definido pela agência brasileira CAPES, responsável pela avaliação oficial). Nesse cenário, também motivado pelo sucesso de muitos grupos de pesquisa devido ao comportamento colaborativo, no próximo capítulo, será apresentada uma nova abordagem para recomendação de colaborações com base em redes sociais acadêmicas.

4 RECOMENDAÇÃO DE COLABORAÇÕES EM REDES SOCIAIS ACADÊMICAS

No capítulo anterior, foi apresentada uma proposta de ranqueamento de grupos de pesquisa com base nas suas colaborações internas, através do uso de métricas especificamente desenvolvidas. Os experimentos apresentados comprovaram a hipótese de que a "conectividade" interna pode ser utilizada como um indicador de qualidade no contexto acadêmico. Dessa forma, mostrou-se que a questão de colaborações e cooperações entre pesquisadores pode ser uma aliada importante na busca por qualidade individual e do grupo como um todo. Sendo assim, como o comportamento colaborativo está diretamente relacionado à qualidade, o surgimento de novas colaborações deve ser estimulado, visando-se um consequente incremento na "qualidade" dos grupos de pesquisa. Neste cenário, recomendar colaborações entre pesquisadores pode ser de grande valia para auxiliar no encontro de novas possibilidades de parcerias.

Dessa forma, a presente tese também explorou a recomendação de colaborações no contexto acadêmico. Neste capítulo, é apresentada uma abordagem proposta para Recomendação de Colaborações em Redes Sociais Acadêmicas. O restante deste capítulo está organizado da seguinte forma. A seção 4.1 apresenta alguns conceitos envolvidos e uma visão geral da abordagem proposta. A seção 4.2 detalha os indicadores definidos e as métricas correspondentes, bem como a função de recomendação propriamente dita. A seção 4.3 apresenta alguns refinamentos para ponderação dos vínculos relacionais entre pesquisadores (refinamentos na métrica de cooperação). Por fim, as seções 4.4 e 4.5 apresentam os experimentos efetuados.

4.1 Conceitos e Visão Geral

Redes Sociais são baseadas na importância dos relacionamentos entre unidades de interação. As unidades de interação das Redes Sociais são conhecidas como *atores* e os relacionamentos entre eles são chamados *vínculos relacionais* (NEWMAN, 2003; WASSERMAN; FAUST, 1994). O "grau" da relação pode ser quantificado por um valor de peso entre dois atores, que visa mensurar a importância do relacionamento existente entre eles. A determinação dos pesos dos relacionamentos entre atores de uma Rede Social é um grande desafio.

No presente trabalho, a Rede Social analisada é uma rede de colaboração científica. De acordo com Newman (2003), que estudou redes de colaboração científica nas quais dois cientistas são considerados conectados se eles têm pelo menos um artigo em coautoria, essa parece ser uma definição razoável de conhecimento científico. Para apresentação da abordagem deste trabalho, uma Rede Social *SN*, especificamente uma rede de coau-

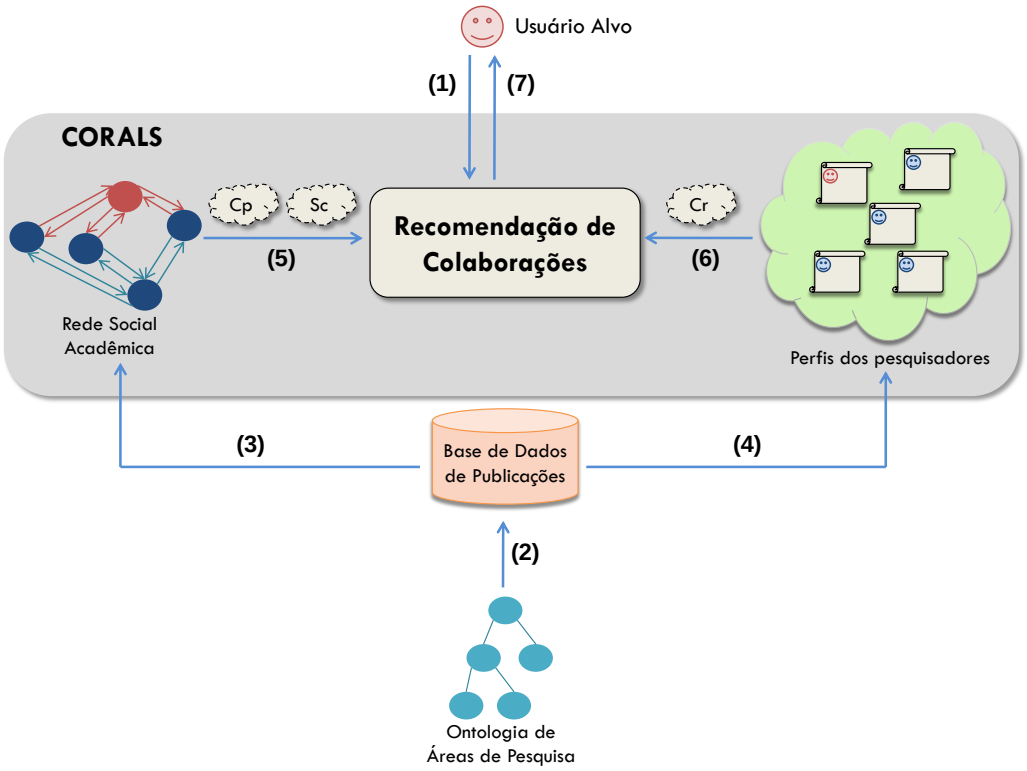


Figura 4.1: Visão geral da **CORALS** (COllaboration Recommender for Academic sociaL networkS): (1) selecionar o usuário alvo; (2) carregar a ontologia de áreas de pesquisa; (3) construir a rede social de colaboração; (4) definir o perfil dos usuários das publicações; (5) calcular a cooperação e a proximidade social; (6) calcular a correlação; e (7) aplicar a função de recomendação e apresentar os resultados ao usuário alvo.

toria, é um par: $SN = (N, E)$ onde N e E são o conjunto de nós (Nodes) e arestas (Edges), respectivamente. Cada aresta $e \in E$ é uma tupla na forma $\langle i, j, w_{i,j} \rangle$, onde a aresta é direcionada de i para j, e w corresponde ao peso afetado pela associação. É importante notar que, neste trabalho, são consideradas redes sociais nas quais todos os nós têm o mesmo tipo e todas as arestas representam um único tipo de relacionamento (em oposição a redes sociais com nós e/ou arestas "heterogêneas", como apresentado, por exemplo, em (ANGELOVA; KASNECI; WEIKUM, 2011)).

Nesta seção, é apresentada uma visão geral da abordagem de recomendação desta tese, nomeada **CORALS** (*COllaboration Recommender for Academic sociaL networkS*), para efetuar a recomendação de colaborações no contexto de Redes Sociais Acadêmicas, conforme ilustrado na Figura 4.1.

O primeiro passo consiste em selecionar o usuário alvo de recomendação (passo 1 na Figura 4.1) e o grupo de indivíduos, ou seja, pesquisadores, que irá compor a Rede Social Acadêmica (SN) a ser construída. Essa flexibilidade permite que um grupo de indivíduos desejado seja selecionado e a abordagem de recomendação pode ser efetuada somente utilizando esse grupo. Por exemplo, na avaliação experimental (seção 4.5), um subconjunto da base de dados da DBLP, composto por pesquisadores de Programas de Pós-graduação em Ciência da Computação brasileiros, foi selecionado. As informações de publicações dos pesquisadores podem ser obtidas de uma Biblioteca Digital como, por exemplo, a DBLP. Se a fonte de informação sobre publicações não contiver as áreas de pesquisa associadas às publicações, um passo de classificação se torna necessário. A

classificação das publicações pode ser feita com a adição de uma Ontologia de Áreas de Pesquisa. Essa classificação é carregada na base de dados de publicações (passo 2). Maiores detalhes sobre esse tipo de ontologia podem ser obtidos na seção 4.5.1.2.

A Rede Social Acadêmica formada por indivíduos selecionados é construída e são atribuídos pesos aos relacionamentos de colaboração (em coautorias) entre eles (passo 3), com o uso das informações contidas na Base de Dados de Publicações (do passo 2). Esses pesos estimam o nível de *cooperação* (*cooperation*-Cp) (discutido na seção 4.2.1), e também são usados na estimativa do nível de *proximidade social* (*social closeness*-Sc) entre pares de indivíduos (discutido na seção 4.2.2).

Além disso, os perfis dos pesquisadores (perfis dos indivíduos da rede de colaboração) são construídos com base na informação disponibilizada sobre suas publicações na Base de Dados de Publicações (passo 4). Os perfis dos pesquisadores são formados pelas áreas de pesquisa nas quais os pesquisadores trabalham. Pesos são associados para representar os graus de atuação dos pesquisadores em cada uma das áreas de pesquisa (grau de importância de cada área para representação do perfil do usuário). A similaridade entre perfis de pesquisadores estima o nível de *correlação* (Cr) entre pares de indivíduos (seção 4.2.3).

Cooperação, *proximidade social* e *correlação* foram os indicadores definidos para serem aplicados na determinação da função de recomendação. Dessa forma, métricas foram definidas para calcular os valores correspondentes dos indicadores. Assim, a função de recomendação pode então ser calculada entre o usuário alvo e os pesquisadores da Rede Social de forma a gerar recomendações (passos 5 e 6). As recomendações sugerem pesquisadores com os quais o usuário alvo pode estabelecer novas colaborações ou com os quais ele deve intensificar as colaborações já existentes (seção 4.2.4). Finalmente, essas recomendações são apresentadas ao usuário alvo de recomendação (passo 7).

Essa abordagem é detalhada nas seções seguintes, onde são propostas as métricas correspondentes aos indicadores (Seções de 4.2.1 até 4.2.3). Além disso, é definida uma nova função de recomendação que combina os três indicadores (cooperação, correlação e proximidade social) a fim de recomendar novas colaborações (Seções 4.2.4). Essa função também apresenta um novo tipo de resultado que é a recomendação de intensificação de colaborações já existentes, o que nunca havia sido feito anteriormente.

4.2 Métricas e Função de Recomendação

Nesta seção, são detalhadas as métricas utilizadas na função de recomendação - *Cooperação* (seção 4.2.1), *Proximidade Social* (seção 4.2.2) e *Correlação* (seção 4.2.3) - e o modo de combinar essas métricas para gerar e ranquear recomendações (seção 4.2.4).

4.2.1 Cooperação

Esta primeira métrica determina um tipo de associação nomeado *Cooperação* (*Cooperation* - Cp), que define a cooperação entre autores em uma rede de colaboração acadêmica. A métrica Cp estima um valor numérico entre 0 e 1 dado pela Equação 4.1.

$$Cp_{i,j} = \frac{p_{i,j}}{p_i} \tag{4.1}$$

onde:

- $Cp_{i,j}$ corresponde ao nível de cooperação com base no relacionamento de coautoria (a cooperação é diferente de acordo com a direção da relação, ou seja, $Cp_{i,j}$ é diferente de $Cp_{j,i}$);

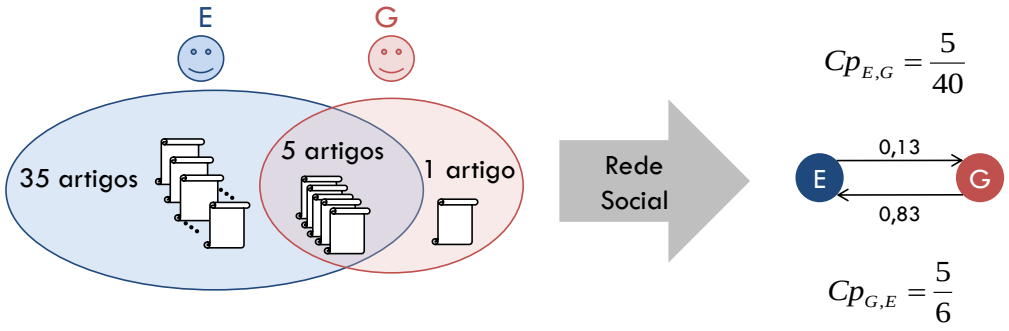


Figura 4.2: Exemplo do cálculo da métrica de *cooperação* entre dois pesquisadores: um orientador E e seu orientando de doutorado G.

- $p_{i,j}$ corresponde ao número de artigos que o autor i tem em coautoria com o autor j;
- p_i corresponde ao número total de artigos do autor i.

Note que quanto maior o peso Cp, maior a importância (relacionada à intensidade) do relacionamento entre o autor j para o autor i. O uso do peso Cp implica que existe um grafo (para representar a rede) com 0 ou 2 arestas entre pares de autores. O peso representa o grau de cooperação em coautorias entre esses autores. Esse peso é uma variante assimétrica do *Coeficiente de Jaccard* e tem sido aplicado no contexto de Redes Sociais (ALEMAN-MEZA et al., 2006; HWANG; WEI; LIAO, 2010; MIKA, 2004).

Por exemplo, considerando o relacionamento entre um doutorando e seu orientador. Normalmente, o doutorando tem a maioria de suas publicações em colaboração com seu orientador. Por outro lado, o orientador provavelmente tem muitas publicações com outros pesquisadores. Então, o peso na direção do orientador para o orientando ($Cp_{orientador,orientando}$) é menor que o peso do orientado para o orientador ($Cp_{orientando,orientador}$). A Figura 4.2 ilustra esse caso considerando E como o orientador e G como o orientando.

Neste exemplo, o número de artigos em coautoria entre E e G é 5. O número total de publicações de E é 40 (5 artigos em coautoria com G e 35 artigos sem a coautoria de G) e o número total de publicações de G é 6 (5 artigos em coautoria com E e somente 1 artigo sem a coautoria de E). O cálculo dos pesos (em ambas as direções) entre E e G estão apresentados a seguir.

$$Cp_{E,G} = \frac{p_{E,G}}{p_E} = \frac{5}{40} \cong 0,13$$

$$Cp_{G,E} = \frac{p_{G,E}}{p_G} = \frac{5}{6} \cong 0,83$$

A Figura 4.3 mostra uma Rede Social completa incluindo os pesquisadores E e G e outros pesquisadores (incluindo outros professores e estudantes). Essa Rede Social também ilustra exemplos do cálculo de outras métricas ao longo desta seção.

4.2.2 Proximidade Social

Tradicionalmente, uma função de recomendação que considera aspectos estruturais de uma Rede Social objetiva estabelecer um nível de *proximidade social* entre seus indivíduos. A determinação da *proximidade social* entre indivíduos frequentemente adota

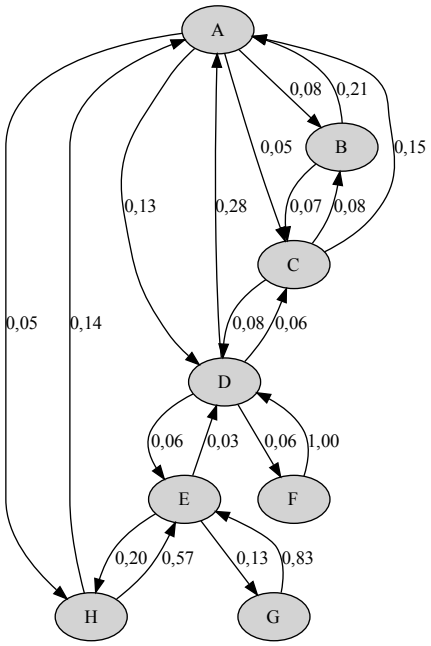


Figura 4.3: Exemplo de uma Rede Social representada por um grafo bi-direcional utilizando os pesos de *cooperação* (*Cp*).

métricas que são originárias da teoria de grafos (discutidas previamente na seção 2.3.2) mas, apenas recentemente, seu potencial para recomendar novas conexões tem sido explorado.

Na abordagem desta tese foi escolhida uma métrica de *proximidade social* (Sc) determinada por uma variante normalizada do método do *menor caminho*. Essa métrica foi escolhida porque é adequada para lidar com uma rede incompleta (e ainda assim terá chances de estimar um nível de proximidade entre pesquisadores). Note que dificilmente se terá acesso a todas as publicações e colaboradores dos pesquisadores, assim, o caso de uma rede incompleta é o mais provável de acontecer. Finalmente, outras abordagens podem ser ainda mais depreciadas nessa situação como, por exemplo, a dos *vizinhos em comum*, na qual um vínculo relacional perdido pode, significantemente, depreciar o resultado final[1]. Além disso, pelo fato de estar-se lidando, muitas vezes, com uma rede incompleta, com o uso do *menor caminho* pode-ser aumentar a possibilidade de conseguir-se estimar algum valor para o indicador de proximidade social.

Especificamente, a ideia principal é de que a proximidade dentro de um grupo apresenta evidências de uma tendência de trabalho colaborativo, através de uma cadeia de conexões interpessoais em uma rede de colaboração. A métrica do *menor caminho* é usada na teoria de grafos para encontrar um caminho entre dois vértices (nós) tal que a soma do peso de suas arestas constituintes seja minimizado. Dessa forma, uma questão importante é como definir um peso para uma aresta entre dois pesquisadores. Note que o peso não deve ser simplesmente 1, porque senão o menor caminho irá considerar somente o número de arestas entre quaisquer dois vértices. Uma solução mais acurada é considerar algo como a cooperação entre dois autores. Especificamente, uma boa definição para o peso pode ser: quanto maior a cooperação, menor o peso; isto é, menor a distância entre os autores. Assim, o peso de uma aresta é definido pela distância d calculada pela

[1]Para maiores detalhes sobre predição de ligações, o trabalho de Liben-Nowell e Kleinberg (2007) apresenta um estudo sobre preditores com base em diferentes métricas de proximidade em grafos.

Equação 4.2.

Em outras palavras, d considera dois custos quando existe uma ligação direta entre dois autores ($Cp \neq 0$): (i) o valor 1 representa o custo da existência de uma ligação, e (ii) um valor em um intervalo $[0, 1)$ representa um custo adicional baseado na métrica de *cooperação*. Nessa situação, o d resultante é um valor numérico no intervalo $[1, 2)$; caso contrário, d é ∞.

$$d_{i,j} = \begin{cases} \infty, & \text{se } (Cp_{i,j} = 0) \\ 1 + (1 - Cp_{i,j}), & \text{caso contrário} \end{cases} \tag{4.2}$$

Até agora, foram definidas métricas para calcular a cooperação entre dois autores e suas respectivas distâncias. Tais distâncias são usadas no cálculo do menor caminho entre dois pesquisadores, que será utilizado na métrica de *proximidade social* definida de acordo com a Equação 4.3.

$$Sc_{i,j} = \begin{cases} 0, & \text{se } (shortest\_path(i,j) = \infty) \\ \frac{(MAX+1) - shortest\_path(i,j)}{MAX}, & \text{caso contrário} \end{cases} \tag{4.3}$$

onde:

- $Sc_{i,j}$ corresponde à métrica de *proximidade social* entre os autores i e j, na direção $i \rightarrow j$ (essa métrica é diferente de acordo com a direção da relação);
- $shortest\_path(i, j)$ corresponde ao *menor caminho* calculado entre os autores i e j na direção $i \rightarrow j$ (utilizando a distância d como peso das relações);
- MAX é uma função que retorna o valor do máximo *menor caminho* encontrado entre dois autores da Rede Social SN (isto é, $MAX = shortest\_\ path(k, l)$, onde k e l é o par de autores cujo *menor caminho* entre eles, na direção $k \rightarrow l$, é o maior de todos os *menores caminhos* entre todos os pares de autores, em todas a direções).

Em outras palavras, quanto maior o valor da métrica $Sc_{i,j}$, em maior "proximidade social" os autores i e j estão. $Sc_{i,j}$ é uma variante normalizada do *menor caminho* representando um valor dentre 0 e 1. Valores próximos de 1 correspondem a pares de autores em maior "proximidade social", enquanto valores próximos a 0 correspondem a pares de autores em menor "proximidade social".

Por exemplo, a Rede Social apresentada na Figura 4.3 foi construída usando a métrica de *cooperação* (pesos Cp). A mesma Rede Social é representada na Figura 4.4 utilizando os pesos d. Esses novos pesos são calculados de acordo com a Equação 4.2 e representam a "distância" entre cada par de autores. Considerando de novo os autores E e G, as seguintes equações apresentam o cálculo dos pesos d em ambas as direções da ligação ($E \rightarrow G$ e $G \rightarrow E$).

$$d_{E,G} = 1 + (1 - Cp_{E,G}) = 1 + (1 - 0,13) = 1,87$$

$$d_{G,E} = 1 + (1 - Cp_{G,E}) = 1 + (1 - 0,83) = 1,17$$

Como as equações mostram, quanto maior a *cooperação* entre autores, menor a "distância" entre eles representada pelo peso d.

Uma vez conhecidos os pesos d, a proximidade social pode ser calculada. Por exemplo, o *menor caminho* entre os autores B e E na direção $B \rightarrow E$ (B é o nó

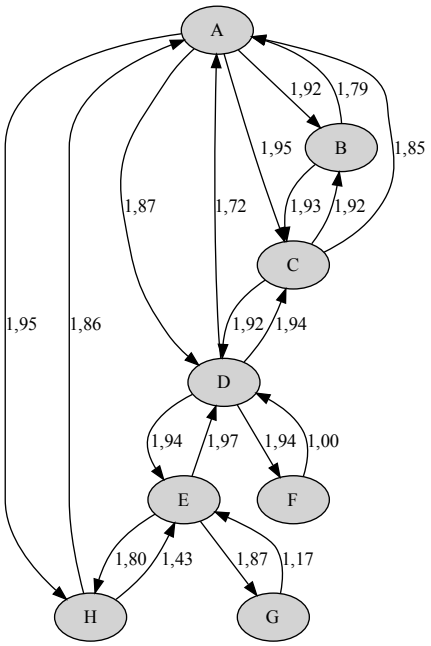


Figura 4.4: Exemplo de uma Rede Social representada por um grafo bi-direcional usando os pesos d.

inicial e E é o nó objetivo) é calculado como apresentado a seguir. Nesse exemplo, o *menor caminho* considerando o número de arestas a serem percorridas é composto por 3 arestas que podem ser $[B \rightarrow A, A \rightarrow H, H \rightarrow E]$, $[B \rightarrow C, C \rightarrow D, D \rightarrow E]$ ou $[B \rightarrow A, A \rightarrow D, D \rightarrow E]$. A primeira opção representa o *menor caminho*, nesse caso, porque considerando os pesos d, a "distância" através de uma cadeia interpessoal de 3 arestas começando no nó B e chegando no nó E passando por A e H, nessa ordem, é a menor. O valor do *menor caminho*, nesse caso, é a soma dos pesos das 3 arestas ($B \rightarrow A$, $A \rightarrow H$ e $H \rightarrow E$), calculado como segue.

$$shortest\_path(B, E) = d_{B,A} + d_{A,H} + d_{H,E} = 1,79 + 1,95 + 1,43 = 5,17$$

Como outro exemplo, considere o *menor caminho* entre os autores B e F na direção $B \rightarrow F$ (B é o nó inicial e F é o nó objetivo). O *menor caminho*, nesse caso, é um valor numérico representando a "distância" através de uma cadeia interpessoal de 3 arestas começando no nó B e alcançando o nó F, passando por A e D, respectivamente. O valor do *menor caminho*, nesse caso, é a soma dos pesos das 3 arestas ($B \rightarrow A$, $A \rightarrow D$ e $D \rightarrow F$), como segue.

$$shortest\_path(B, F) = d_{B,A} + d_{A,D} + d_{D,F} = 1,79 + 1,87 + 1,94 = 5,6$$

Nesses dois exemplos, se os pesos d não tivessem sido utilizados, o resultado do *menor caminho* seria igual a 3 (3 arestas). Entretanto, utilizando os pesos d, o resultado do *menor caminho* é diferente (ou seja, 5,17 e 5,6). Essa diferença possibilita ranquear autores que estejam distantes a um mesmo número de "arestas", mas que tenham diferentes níveis de cooperação através da cadeia de relacionamentos interpessoais.

Finalmente, um exemplo de cálculo da métrica de *proximidade social* é apresentado como segue. A métrica é calculada entre os pares $B \rightarrow E$ e $B \rightarrow F$. Nesse caso, o

autor B está em maior "proximidade social" com o autor E do que com o autor F, porque:

$$Sc_{B,E} = \frac{(MAX+1)-shortest\_path(B,E)}{MAX} = \frac{(7{,}04+1)-5{,}17}{7{,}04} \cong 0,41$$

$$Sc_{B,F} = \frac{(MAX+1)-shortest\_path(B,F)}{MAX} = \frac{(7{,}04+1)-5{,}6}{7{,}04} \cong 0,35$$

Nesse exemplo, MAX representa o máximo valor de *menor caminho* encontrado entre B e G (na direção $B \rightarrow G$); isto é, este valor é a soma dos pesos das arestas ao longo do *menor caminho*, conforme apresentado abaixo.

$MAX = shortest\_path(B,G) = d_{B,A} + d_{A,H} + d_{H,E} + d_{E,G} = 1,79 + 1,95 + 1,43 + 1,87 = 7,04$

4.2.3 Correlação

Para uma recomendação de colaborações mais acurada, a abordagem desta tese também considera uma métrica que determina a correlação entre pesquisadores. Essa correlação considera as diferentes áreas de pesquisa nas quais os pesquisadores atuam e é uma importante faceta no contexto acadêmico. Para esse propósito, são tomadas duas ações: (i) é construído um perfil de atuação para cada pesquisador considerando as áreas de pesquisa nas quais eles atuam; e (ii) são associados pesos para representar o grau de atuação dos pesquisadores nas áreas correspondentes, em relação ao seu perfil global. A correlação é calculada entre os perfis de atuação dos pesquisadores.

A Equação 4.4 calcula o peso de cada área de pesquisa que irá compor o perfil de atuação do pesquisador, chamado $R_{i,k}$.

$$R_{i,k} = \frac{p_{i,k}}{p_i} \tag{4.4}$$

onde $p_{i,k}$ corresponde ao número de artigos que um autor i publicou na área de pesquisa k, e p_i ao número total de artigos de i. Então, cada área k tem um peso correspondente que indica "quantos" dos artigos do autor foram publicados na área k.

Cada peso, por área de pesquisa, estima o grau de importância da área para definição do perfil do pesquisador (com base nas publicações apresentadas na Base de Dados de Publicações considerada). Os pesos são usados no cálculo da correlação entre pares de autores e o Modelo de Espaço Vetorial (*Vector Space Model - VSM*) (SALTON; BUCKLEY, 1988) é utilizado para efetuar essa computação. Tradicionalmente, o VSM define um espaço n-dimensional usado para representar n termos de indexação em um processo de Recuperação de Informações. No presente caso, n corresponde ao número de áreas de pesquisa distintas. Cada perfil de autor é representado por um vetor de áreas de pesquisa, e os pesos representam as coordenadas do vetor na correspondente dimensão. Com base no VSM, a similaridade é calculada entre pares de autores e as dimensões correspondem às áreas de pesquisa dos autores. O peso associado a cada área de pesquisa permite distinção entre as áreas de pesquisa, sendo calculado de acordo com a importância de cada área para cada autor pelo uso da Equação 4.4. Esses pesos variam continuamente entre 0 e 1. Valores próximos a 1 correspondem a áreas de pesquisa mais importantes para o autor, enquanto que valores próximos a 0 correspondem a áreas de pesquisa de menor importância para o autor. O princípio do VSM é baseado na correlação inversa entre a distância entre vetores no espaço e a similaridade da informação (nesse caso os perfis dos

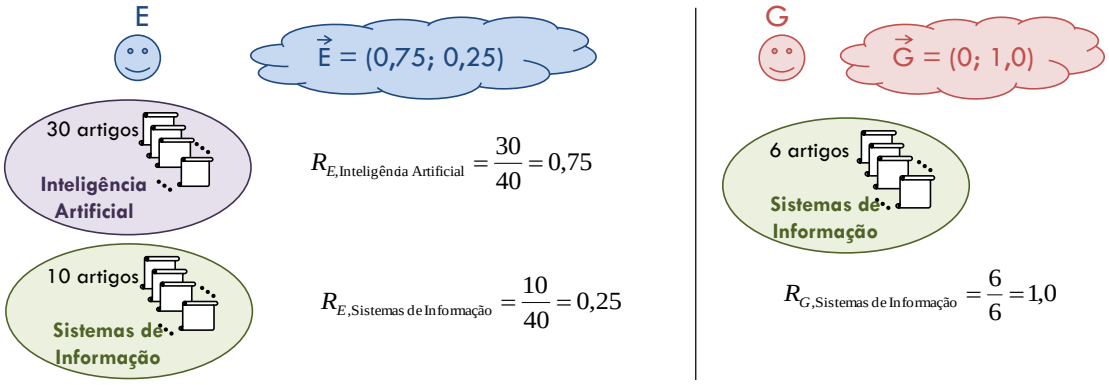


Figura 4.5: Exemplo da determinação de perfis dos usuários.

autores) que eles representam. Para calcular o escore de similaridade, que irá representar a correlação entre dois autores, a Equação 4.5 utiliza o cosseno. O valor resultante indica o grau de correlação (Cr) entre dois perfis de autores (i e j), onde $R_{i,k}$ representa o peso de cada área de pesquisa que compõe o perfil do usuário, e n representa o número total de áreas de pesquisa.

$$Cr_{i,j} = \frac{\sum_{k=1}^{n} R_{i,k}.R_{j,k}}{\sqrt{\sum_{k=1}^{n} (R_{i,k})^2 . \sum_{k=1}^{n} (R_{j,k})^2}} \tag{4.5}$$

Por exemplo, a Figura 4.5 ilustra os perfis de atuação de dois pesquisadores E (orientador) e G (orientando). Nessa representação, está cada área de pesquisa que o pesquisador atua, seguida pelo peso associado representando o grau de atuação do pesquisador na área correspondente (calculado pela Equação 4.4). A situação é a seguinte: (i) pesquisador E atua em duas áreas de pesquisa, sendo 75% dos seus artigos (30 dos seus 40 artigos) em Inteligência Artificial e 25% (10 dos seus 40 artigos) em Sistemas de Informação; e (ii) o pesquisador G atua em uma única área de pesquisa, representando 100% dos seus artigos (6 dos seus 6 artigos) em Sistemas de Informação. Os vetores representando os perfis dos usuários estão representados na Figura 4.5 como $\vec{E}$ e $\vec{G}$. Nesse exemplo, uma das dimensões corresponde à Inteligência Artificial e outra dimensão corresponde a Sistemas de Informação. O autor G não trabalha em Inteligência Artificial, então o peso do seu vetor nessa direção é zero (0). Nesse exemplo, os autores E e G têm a área de pesquisa de Sistemas de Informação em comum.

A Figura 4.6 representa um espaço bidimensional porque o número de áreas distintas nesse exemplo é dois. Nessa figura, cada perfil de autor é representado por vetores de áreas de pesquisa ($\vec{E}$ e $\vec{G}$), e os pesos representam as coordenadas dos vetores na dimensão correspondente (importância da área de pesquisa para a representação do perfil do usuário).

Para calcular o escore de similaridade, que representa a correlação entre os dois autores, a Equação 4.5 é utilizada (que corresponde ao cosseno do ângulo θ representado na Figura 4.6). O cálculo de $Cr_{E,G}$ é apresentado abaixo e o valor resultante (0,3162) representa a *correlação* entre os pesquisadores E e G.

$$Cr_{E,G} = \frac{0{,}75*0+0{,}25*1{,}0}{\sqrt{(0^2+1{,}0^2)*(0{,}75^2+0{,}25^2)}} = 0,3162$$

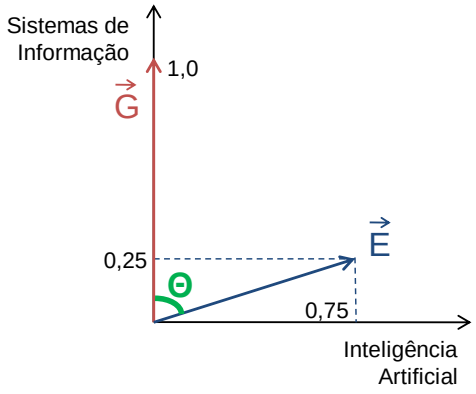


Figura 4.6: Exemplo da representação de perfis dos usuários em um espaço bidimensional.

4.2.4 Recomendação

Nas seções anteriores, foram apresentadas três métricas que irão agora compor a função de recomendação proposta. Dados dois autores i e j:

- a métrica de cooperação quantifica a proporção de artigos de i em coautoria com j (em relação ao total de artigos de i);
- a métrica de proximidade social estabelece o nível de proximidade social quando os autores i e j não têm nenhum artigo em coautoria;
- e a métrica de correlação quantifica o quanto i e j têm publicado em áreas de pesquisa em comum.

Uma das características que distinguem a abordagem desta tese é a possibilidade de recomendar duas diferentes ações: (i) iniciar colaboração quando dois autores não têm artigos em coautoria; e (ii) intensificar a colaboração quando eles já foram coautores juntos. Dessa forma, a função de recomendação deste trabalho precisa considerar ambos os casos.

Para o primeiro caso (i e j ainda não são coautores), a *correlação* e a *proximidade social* são combinadas para estabelecer uma métrica única, nomeada Cr_Sc, utilizando a Equação 4.6.

$$Cr\_Sc_{i,j} = \frac{w_{Cr}.Cr_{i,j} + w_{Sc}.Sc_{i,j}}{w_{Cr} + w_{Sc}} \tag{4.6}$$

onde, a métrica Cr_Sc entre dois autores i e j é calculada por uma média ponderada. Os pesos w_{Cr} e w_{Sc} determinam, respectivamente, a importância das métricas Cr e Sc para o valor resultante. A métrica única Cr_Sc é usada como um escore para ranquear as recomendações de colaborações que serão geradas. Os pesos w_{Cr} e w_{Sc} podem ser usados para enfatizar uma das métricas combinadas. Então, o ranking das recomendações pode ser ajustado de acordo com as seguintes preferências: (i) escore de recomendação enfatiza mais a correlação entre dois autores (quando valores de pesos são definidos como $w_{Cr} > w_{Sc}$), (ii) escore de recomendação enfatiza mais a proximidade social entre autores

($w_{Cr} < w_{Sc}$), ou (iii) escore de recomendação considera igualmente a correlação e a proximidade social entre autores ($w_{Cr} = w_{Sc}$).

Para o segundo caso (i e j já foram coautores juntos), a *cooperação* e a *correlação* são combinadas a fim de definir o escore final. Tal escore é usado para definir o ranking final de recomendação.

O relacionamento entre os indicadores de *cooperação*, *proximidade social* e *correlação* para cada par de autores estabelece a necessidade (ou não) de ter-se uma maior interação de pesquisa entre eles. Para esse tipo de análise, foram estabelecidos graus para representar os diferentes intervalos de valores que são possíveis de serem obtidos para as métricas correspondentes a esses indicadores. Os graus são "high" (alto), "medium" (médio) e "low" (baixo) e podem corresponder a uma escala linear. Por exemplo, os graus "low", "medium" e "high" correspondem, respectivamente, aos valores v nos intervalos $\{v \in \Re | 0.0 < v \leq 1/3\}$, $\{v \in \Re | 1/3 < v \leq 2/3\}$ e $\{v \in \Re | 2/3 < v \leq 1.0\}$. É importante perceber que tais intervalos podem ser otimizados de acordo com o *feedback* do usuário, que será explorado em trabalhos futuros. Além disso, esses graus podem ser futuramente refinados e seus respectivos intervalos, por exemplo, podem ser diferenciados para cada usuário, levando em consideração o valor de peso máximo e mínimo de suas relações pré-existentes.

Com base nessas análises, ações podem ser estabelecidas para indicar possíveis recomendações de colaborações entre pares de autores. A combinação entre os graus dos indicadores e as ações correspondentes que incluem "Initiate_Collaboration" (Iniciar colaboração), "Intensify_Collaboration" (Intensificar colaboração) e "Not_Recommend" (Não recomendar) pode ser observada na Equação 4.7.

$$Action_{i,j} = \begin{cases} \text{Initiate\_Collaboration}, & \text{se} \quad (Cp_{i,j} = 0) \wedge \\ & \quad\ (Cr\_Sc_{i,j} > threshold) \\ \text{Intensify\_Collaboration}, & \text{se} \quad (Cp_{i,j} \in low) \wedge \\ & \quad\ ((Cr_{i,j} \in medium) \vee \\ & \quad\ (Cr_{i,j} \in high)) \\ \text{Not\_Recommend}, & \quad\ \text{caso contrário} \end{cases} \tag{4.7}$$

A Equação 4.7 indica que uma cooperação zero e um Cr_Sc não-zero (maior que um valor de limiar - ou *threshold*, que pode corresponder a um dos valores mínimos dos intervalos correspondentes aos graus) caracteriza uma possível recomendação (correlação existente entre os autores e/ou eles estão em algum nível de proximidade social). Além disso, pares de autores com cooperação baixa ("low"), mas com correlação média ou alta ("medium" ou "high") são recomendados a intensificar suas cooperações. Os outros casos não caracterizam uma ação de recomendação. Por exemplo, pares de autores com cooperação alta ou média ("high" ou "medium") que têm baixos graus de correlação não precisam ser recomendados a cooperar, uma vez que a cooperação entre os autores já está acontecendo de uma forma mais intensa do que a correlação entre seus perfis sugere.

Uma vez que as ações foram estabelecidas, é necessário definir os escores para as possíveis recomendações, de forma que se possa ordená-las para geração do ranking final de recomendações. Com essa finalidade, é definido o escore de recomendação de acordo com a Equação 4.8.

$$Score_{i,j} = \begin{cases} Cr\_Sc_{i,j}, & \text{se } (Action_{i,j} = \text{Initiate\_Collaboration}) \\ \frac{Cp_{i,j}}{Cr_{i,j}}, & \text{se } (Action_{i,j} = \text{Intensify\_Collaboration}) \\ 0, & \text{caso contrário} \end{cases} \quad (4.8)$$

Como apresentado anteriormente, se a ação de recomendação é para Iniciar colaboração ("Initiate_Collaboration"), o escore de recomendação é calculado pela métrica Cr_Sc. Se a ação de recomendação é para Intensificar colaboração ("Intensify_Collaboration"), a razão entre cooperação e correlação é apresentada. Além disso, as recomendações são apresentadas em uma lista ranqueada: na lista de Iniciar colaboração ("Initiate_collaboration"), os pesquisadores recomendados são ordenados em ordem decrescente da métrica Cr_Sc, uma vez que quanto maior o valor da métrica, maior a possibilidade de a colaboração ser interessante para o usuário alvo de recomendação; e, na lista de Intensificação de colaboração ("Intensify_collaboration"), os pesquisadores recomendados são apresentados em ordem crescente da razão entre cooperação e correlação, uma vez que as recomendações são apresentadas em ordem de necessidade de intensificação.

4.3 Consideração de Aspectos Temporais para refinamento na ponderação de vínculos relacionais

Um aspecto importante a ser considerado em relação às colaborações prévias dos pesquisadores diz respeito ao ano em que estas ocorreram. Sendo assim, a consideração de aspectos temporais também é um fator que deve ser explorado. Para tanto, nesta seção são apresentadas propostas de métricas refinadas para ponderação dos vínculos relacionais (refinamento da métrica de *Cooperação*), especificamente, com relação à consideração de aspectos temporais.

Em uma rede de coautoria, os atores são os pesquisadores e os vínculos relacionais são as colaborações de pesquisa entre pares de pesquisadores. A maioria das abordagens de recomendação não considera os ricos aspectos envolvidos nos vínculos relacionais. Além disso, diferentes aspectos podem ser considerados para definição de pesos de vínculos relacionais, e o mais comum está relacionado à quantidade de publicações. No entanto, nesta tese são propostas alternativas para considerar aspectos temporais em métodos de ponderação de vínculos relacionais.

Determinação de pesos para vínculos relacionais. O processo para determinação de um peso p para cada vínculo relacional de uma Rede Social Acadêmica usualmente considera a quantidade de publicações de cada par de pesquisadores $< i, j >$, como dado pela Equação 4.9.

$$p_{ij} = \frac{n_{ij}}{n_i} \quad (4.9)$$

onde n_{ij} corresponde ao número de artigos em comum entre $< i, j >$ e n_i ao número de artigos do autor i. Esses pesos não simétricos, uma vez que p_{ij} é diferente de p_{ji} quando n_i é diferente de n_j. Essa métrica é uma variante assimétrica do coeficiente de *Jaccard*

que já tem sido aplicado no contexto de Redes Sociais (ALEMAN-MEZA et al., 2006; HWANG; WEI; LIAO, 2010), quer como uma variante simétrica ou assimétrica.

Novos pesos considerando aspectos temporais. Este é o primeiro trabalho que considera a influência de aspectos temporais para determinar os pesos relacionais em uma Rede Social que é usada como base para recomendação de colaborações acadêmicas (não foram encontrados outros trabalhos correlatos). Os aspectos temporais são relacionados ao ano de publicação dos artigos que são considerados para a construção da Rede Social. Os valores de ano são usados para inferir quão recente são as colaborações entre os pesquisadores. Desse modo, são definidos maiores valores para relacionamentos que foram ativados mais recentemente.

A Equação 4.10 apresenta o fator de aspecto temporal que depois pode ser considerado para estabelecimento de outros pesos.

$$t_k = \begin{cases} \frac{w-(y_r-y_k)}{w}, & \text{se } (y_r - y_k) < w \\ t_{min}, & \text{caso contrário} \end{cases} \tag{4.10}$$

onde w é o intervalo de tempo para ser depreciado proporcionalmente (do ano de publicação mais recente para o mais antigo, que é depreciado), y_r denota o ano de publicação mais recente considerado na construção da Rede Social, y_k denota o ano de publicação para o qual o fator temporal é calculado, e t_{min} é o valor mínimo de t_k que pode ser gerado e aplicado para anos cuja diferença para y_r é maior ou igual a w.

A seguir, são propostos dois diferentes pesos que consideram tais aspectos temporais. O primeiro peso, chamado Tr, aplica o fator temporal para o ano de publicação mais recente entre dois autores. Esse peso objetiva definir maiores valores a relacionamentos que foram ativados recentemente. A Equação 4.11 apresenta sua fórmula.

$$Tr_{ij} = p_{ij}.t_{y_{ij}} \tag{4.11}$$

onde p_{ij} representa o peso considerando a quantidade de publicações e $t_{y_{ij}}$ indica o ano de publicação em coautoria mais recente entre o par de pesquisadores $< i, j >$.

O segundo peso, chamado T_g, aplica o fator temporal considerando todos os anos das publicações dos pesquisadores. Onde, p_{ij} é modificado para considerar o fator temporal para calcular ambos: numerador e denominador. O ano de publicação é considerado na equação para todas as publicações. Esse peso visa definir altos valores para relacionamentos que foram ativados mais vezes recentemente e é dado pela Equação 4.12.

$$Tg_{ij} = \frac{\sum_{k=1}^{n_{ij}} t_{y_k}}{\sum_{x=1}^{n_i} t_{y_x}} \tag{4.12}$$

onde n_{ij} denota o número de artigos em comum entre $< i, j >$, n_i o número de artigos do autor i, e t o fator temporal sendo avaliado de acordo com cada ano de publicação sendo considerado.

Como um exemplo, considere a rede apresentada na Figura 4.7(a) para o usuário 21, que é também o usuário alvo da função de recomendação. Por simplicidade, a Figura 4.7(a) mostra os pesos somente em uma direção com valores utilizando os métodos p (Equação 4.9) e Tr (Equação 4.11). Esses pesos entre os pares de pesquisadores $< 21, 33 >$, $< 21, 43 >$ e $< 33, 43 >$ são os mesmos utilizando ambos os métodos. Esse caso indica que tais pares tiveram coautorias de artigos no ano mais recente considerado

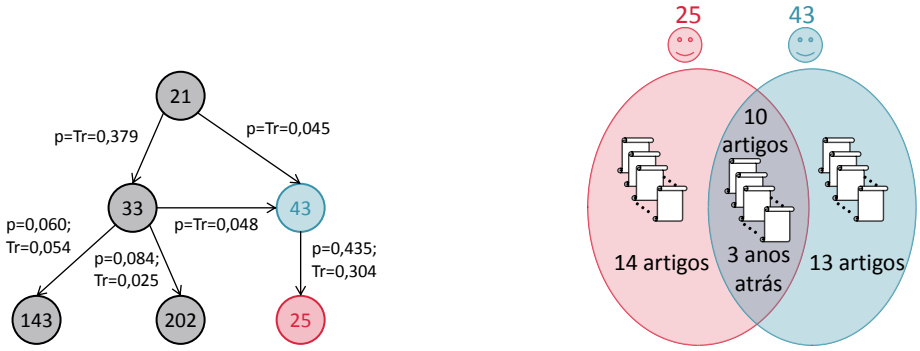


Figura 4.7: Exemplo da definição de pesos em uma Rede Social: (a) rede parcial e (b) publicações em comum.

(parâmetro y_r do fator temporal). Nos outros casos, os pesos sofrem depreciação temporal. Por exemplo, o comportamento de colaboração para o par $< 43, 25 >$ é ilustrado na Figura 4.7(b). Note que, para esse par, existe uma redução comparando o valor dos pesos p e Tr. Especificamente, o valor de $p_{43,25}$ é aproximadamente 0,435 (total de artigos em coautoria entre esses autores, dividido pelo número total de publicações do autor 43, ou seja, $10/23$). Entretanto, para calcular Tr, o valor de p é depreciado usando $t_{y_{43,25}}$. Nesse exemplo, o parâmetro w (representando a janela de tempo de depreciação) usado é igual a 10. A diferença entre o ano mais recente e o ano da mais recente publicação em coautoria pelo par $< 43, 25 >$ é de 3 anos atrás. O fator temporal obtido corresponde a $0,7$ (ou seja, $(10 - 3)/10$). Então, o valor final de tr é aproximadamente 0,304 (ou seja, $p_{43,25} * t_{y_{43,25}} = 0,435 * 0,7$).

Esses pesos, considerando aspectos temporais, são usados para encontrar recomendações para os usuários. Então, um método de escore deve ser aplicado, considerando esses pesos, para ranquear os resultados. Nesse caso, como os pesos representam a "proximidade" entre usuários, cada peso final a ser usado pelo método de menor caminho deve ser calculado como: ∞, se o valor do peso representando a "proximidade" for igual a 0 (pesquisadores não conectados); 1 menos o valor do peso de "proximidade", caso contrário. Para ranquear os resultados de recomendação de novas colaborações, pesquisadores diretamente conectados são descartados (não recomendados), e a saída restante do método de escore de menor caminho deve ser ordenada crescentemente, quanto menor o valor do resultado em maior proximidade social os pesquisadores estão.

4.4 Estudo de caso sobre Intensificação de Colaborações

Nesta seção, é apresentado um estudo de caso preliminar para mostrar a validade e aplicabilidade do tipo de recomendação de "Intensificação de Colaborações" (apresentado previamente na seção 4.2.4). Nesse estudo de caso, é analisada uma rede composta por 27 pesquisadores envolvidos no InWeb, Instituto Nacional de Ciência e Tecnologia para Web[2]. Esse projeto começou em 2008 e todos os seus pesquisadores são membros do corpo docente (incluindo professores titulares, associados, adjuntos e assistentes) nas instituições de ensino e pesquisa brasileiras participantes (UFMG, UFRGS, UFAM e CEFET-MG) com programas de Pós-graduação em Ciência da Computação. Detalhes so-

[2]InWeb: http://inweb.org.br

bre as instituições associadas podem ser obtidos na Tabela 3.1 (apresentada previamente na seção 3.3.1). Aqui, o conjunto de dados sobre as publicações dos pesquisadores, utilizado para construção da rede do InWeb, foi obtido pela equipe do CiênciaBrasil - Portal de Ciência & Tecnologia[3], que realizou uma coleta dos dados dos currículos Lattes dos pesquisadores em 29 de novembro de 2010. Maiores detalhes sobre essa coleta e as técnicas empregadas para desambiguação de nomes e publicações podem ser obtidos em (LAENDER et al., 2011). Além disso, para a identificação das áreas de pesquisa associadas às publicações dos pesquisadores, foi utilizada a classificação da ACM (*Association for Computing Machinery*)[4] disponibilizada para os artigos dos pesquisadores que se encontram indexados na biblioteca digital da ACM[5] (especificamente, utilizou-se o segundo nível dessa classificação). Tal classificação foi obtida através de uma extração de dados da referida biblioteca ocorrida em 09 de dezembro de 2011. Não se utilizou uma ontologia para fazer a classificação automatizada, visando-se trabalhar com uma maior acurácia na identificação das áreas de forma que o valor de Correlação obtido, que depende diretamente da identificação adequada das áreas de pesquisa, seja o mais fiel possível à realidade. Apesar de ter-se um conjunto mais restrito de publicações classificadas, estas estarão corretamente identificadas, já que a classificação apresentada na biblioteca da ACM foi indicada pelos próprios autores no momento da publicação.

A fim de avaliar a evolução das relações entre pesquisadores, através de coautorias, foram escolhidos dois diferentes intervalos de tempo. Esses intervalos são relacionados aos anos de publicação dos artigos considerados para determinar as relações de coautoria. No primeiro intervalo, são consideradas as publicações dos autores até o ano de 2007 (antes e incluindo). Esse intervalo de tempo foi aplicado para construir os perfis dos pesquisadores e a rede social usados para calcular as métricas de Cooperação e Correlação (usadas na função de recomendação, apresentadas previamente nas seções 4.2.1 e 4.2.2). No segundo intervalo, são consideradas todas as publicações dos autores até o ano de 2010, visando representar a situação "atual" (na época que os dados foram coletados) das colaborações entre os pesquisadores. Como o projeto começou em 2008, então o primeiro intervalo considerado reflete as colaborações entre os pesquisadores, antes do início do projeto InWeb, enquanto que o segundo mostra as colaborações dos pesquisadores até três anos depois do início do projeto, ou seja, durante o seu desenvolvimento. Esse conjunto de dados foi escolhido porque se pode considerar que os pesquisadores foram "recomendados" a cooperarem com a criação desse projeto. Dessa forma, foi estudada a evolução ocorrida na rede de colaboração, comparando as relações presentes no primeiro e no segundo intervalos. A Figura 4.8 mostra graficamente esta evolução. Os números dentro dos nós especificam os pesquisadores: o primeiro número (#Id) identifica a instituição de afiliação dos pesquisadores de acordo com a Tabela 3.1, e o segundo número identifica cada pesquisador dentro da instituição. Os pesquisadores que têm pelo menos um artigo em coautoria estão conectados pelas arestas: linhas contínuas em cinza representam conexões que não foram intensificadas durante o projeto InWeb; linhas contínuas em preto representam as conexões intensificadas durante o desenvolvimento do projeto (de 2008 até 2010); e as linhas tracejadas representam as novas conexões entre pesquisadores iniciadas durante o projeto até o ano de 2010.

Foi identificado um total de 67 colaborações existentes entre os pesquisadores até 2007, sendo que, dessas, 38 foram intensificadas no período posterior, identificadas pela

[3]Portal CiênciaBrasil: http://www.pbct.inweb.org.br/pbct/

[4]ACM Computing Classification: http://www.acm.org/about/class/ccs98-html

[5]ACM Digital Library: http://dl.acm.org/

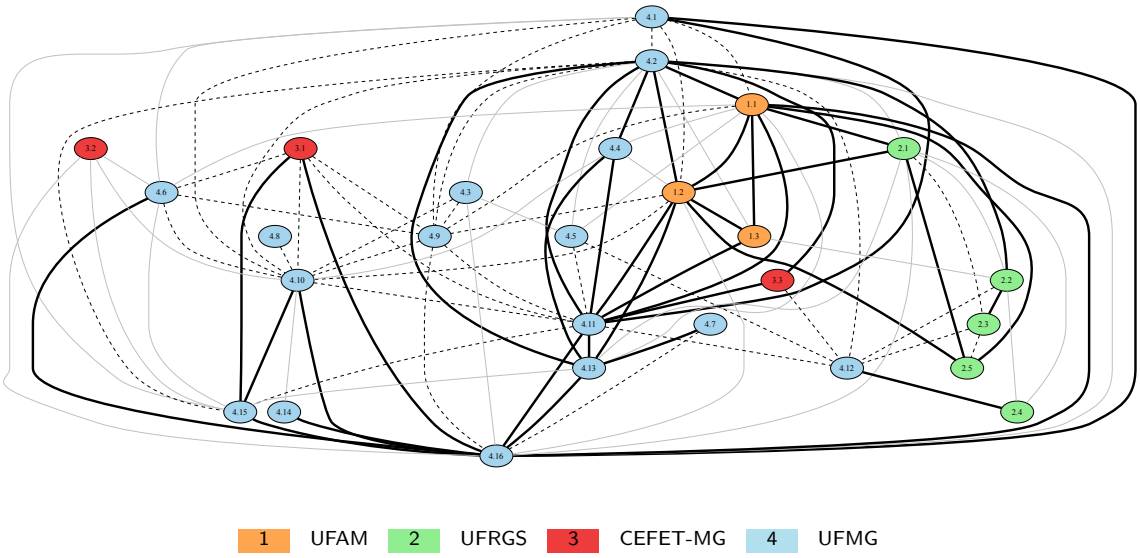


Figura 4.8: Comparativo entre a Rede Social do InWeb antes do início do projeto e durante seu desenvolvimento (com base em dados dos currículos Lattes dos pesquisadores): linhas cinzas para conexões não intensificadas, linhas pretas para conexões intensificadas, e linhas tracejadas para novas conexões.

presença de novas publicações em coautoria após 2007. Assim, a recomendação de "Intensificação de Colaborações" deveria ser calculada com dados até 2007 e os resultados ideais dessa recomendação poderiam ser considerados as intensificações ocorridas ao longo do desenvolvimento do projeto, porque caracterizam as relações que realmente ocorreram após uma "recomendação". Na Figura 4.9, é apresentado um gráfico que mostra a situação dos valores de Correlação e Cooperação calculados entre pares de pesquisadores que já cooperavam na rede do InWeb em 2007 (métrica de Cooperação não zero), com base nos dados de publicação do primeiro intervalo de tempo. Para simplificar a visualização, o menor valor de Cooperação (que ocorre em um dos sentidos da ligação entre pares de pesquisadores) é representado no eixo y. Além disso, o valor de Correlação entre ambos é representado no eixo x. A tendência que se observa nesse gráfico é de que quanto maior a Correlação, maior o número de relações que foram intensificadas (maior quantidade de asteriscos vermelhos). Já, quanto menor a Correlação, menor o número de intensificações (menor a quantidade de asteriscos vermelhos). Dessa forma, uma função de recomendação que considere a relação entre Correlação e Cooperação pode ser de grande valia. Cabe observar que houve resultados de Correlação zero entre pesquisadores que já cooperaram previamente, isto ocorreu porque estes não tiveram nenhuma publicação em coautoria que pudesse ter identificada sua correspondente área de pesquisa no portal da ACM e também não houve nenhuma área de pesquisa em comum dentre as publicações restantes de ambos. A ideia da função de recomendação com tipo de resultado de "Intensificação de Colaborações" parte do pressuposto que pesquisadores que já cooperam em níveis elevados nem precisam ser recomendados para continuarem essa colaboração (na rede do estudo de caso, realmente as colaborações entre pesquisadores com valores de Cooperação mais elevados, continuaram na maioria dos casos). Além disso, mesmo em colaborações menos estabelecidas, em que pesquisadores ainda estão colaborando pouco (valores de Cooperação mais baixos), mas que possuíam uma Correlação mais elevada, também intensificaram essas colaborações. Esses resultados obtidos mostram indícios de que as proposições adotadas para a função de recomendação

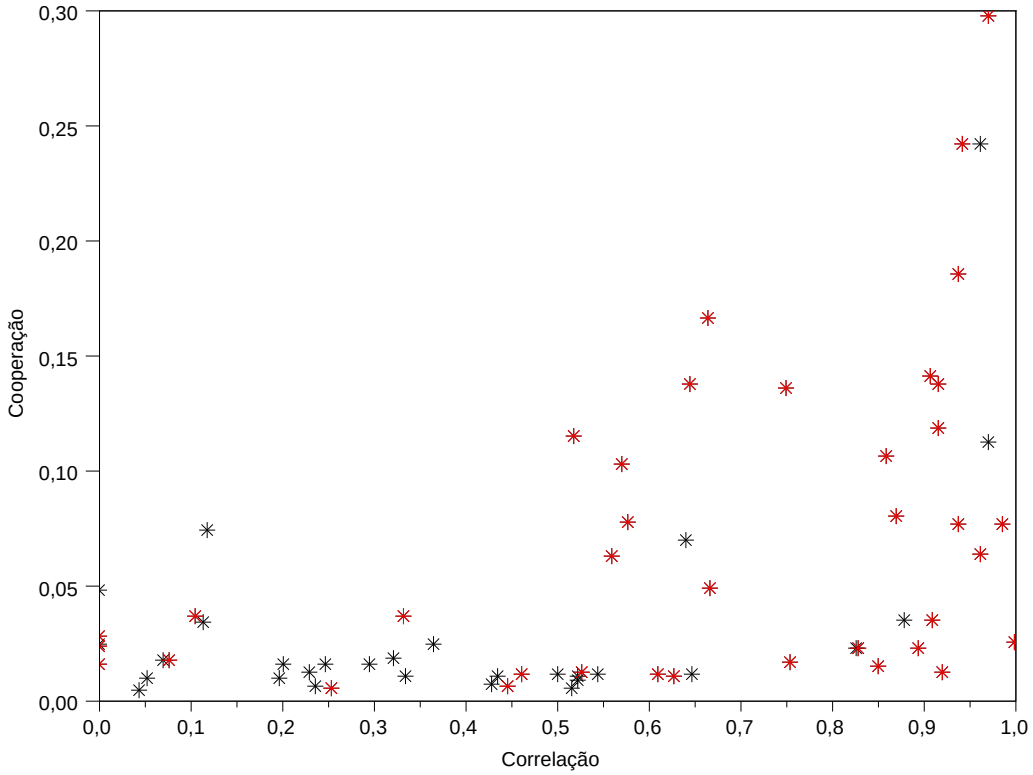


Figura 4.9: Gráfico de Cooperação *versus* Correlação entre pares de pesquisadores do Projeto InWeb que já possuíam alguma relação de coautoria iniciada até 2007. Asteriscos em vermelho indicam relações que foram intensificadas no período de 2008-2010. Asteriscos em preto indicam relações que não sofreram alteração.

com tipo de resultado de "Intensificação de Colaborações" estão refletindo a realidade das intensificações de colaborações quando ocorre uma recomendação.

4.5 Avaliação Experimental

Nesta seção, são apresentados experimentos relativos à abordagem de recomendação proposta nesta tese. A seguir, são detalhadas as configurações efetuadas para realização dos experimentos (seção 4.5.1), e apresentados os dois grandes grupos de experimentos efetuados: o primeiro relativo à abordagem de recomendação como um todo (seção 4.5.2) e o segundo relativo ao impacto dos refinamentos na ponderação de vínculos relacionais pela consideração de aspectos temporais (seção 4.5.3).

4.5.1 Configurações dos experimentos

Esta seção descreve em detalhes as definições e configurações utilizadas na avaliação experimental apresentada nas próximas seções (seções 4.5.2 e 4.5.3). Especificamente, a seção 4.5.1.1 discute o conjunto de dados dos quais são coletadas as informações sobre os artigos dos autores; a seção 4.5.1.2 introduz como foram definidas as áreas de pesquisa associadas aos artigos; e a seção 4.5.1.3 apresenta o conjunto de pesquisadores considerados como autores.

4.5.1.1 Conjunto de dados de pesquisadores

Foi implementada uma ferramenta para automaticamente gerar uma rede social. Tal rede foi construída utilizando informações sobre autores provenientes da DBLP[6] em 03 de agosto de 2010. É importante observar que essa biblioteca exporta seus dados como um documento XML. Ao invés de utilizar todo o conjunto de dados, foram extraídos apenas os artigos publicados em anais de conferências e em periódicos (isto é, elementos *inproceedings* e *article* do documento XML). Este processo de coleta totalizou 831.994 autores; 844.816 artigos em anais de conferências e 561.215 artigos em periódicos. Tal subconjunto foi escolhido porque essa informação é suficiente para identificar os relacionamentos de coautoria entre autores e, consequentemente, as principais colaborações de pesquisa entre eles. Estudos recentes (LAENDER et al., 2008; REITZ; HOFFMANN, 2010) discutem que a cobertura das subáreas da Ciência da Computação pela DBLP tenha atingido valores de aproximadamente 67%, cobrindo acima disto até 96% de algumas subáreas. Sendo assim, a DBLP tem sido amplamente aplicada para obter publicações da área da Ciência da Computação, tendo indexados 1,8 milhões de artigos em Dezembro de 2011. Porém, não se descarta a possibilidade de que alguns resultados de exceção possam ter sido motivados pela limitada cobertura para alguma subárea em específico.

Para escolha da(s) fonte(s) de dados, outra importante consideração diz respeito à qualidade desses dados. Alguns exemplos de problemas que podem ocorrer e que precisam de soluções relacionadas à limpeza de dados (*data cleaning*) (RAHM; DO, 2000) são os seguintes: podem ocorrer desde erros de grafia durante a entrada de dados, informações em falta ou dados inválidos até problemas em decorrência, principalmente, da utilização de múltiplas fontes que precisam ser integradas, como dados redundantes e com diferentes representações/variações. Tendo consciência dessas dificuldades, uma das razões da escolha da DBLP também diz respeito justamente à presença de qualidade nos dados (metadados para indexação) nesta armazenados.

Mesmo na DBLP, por se tratar de um vasto conjunto de dados, a desambiguação dos nomes dos autores pode ser um problema significativo. Por exemplo, alguns autores têm diferentes representações de seus nomes em bibliotecas digitais, ou seja, existem diferentes formas utilizadas para referenciar a mesma pessoa. A desambiguação de nomes pode ser automatizada, empregando técnicas especificamente desenvolvidas para resolver esse problema, tais como (COTA et al., 2010; HAN et al., 2004; SONG et al., 2007). Para esta avaliação experimental, o processo de desambiguação dos nomes dos autores foi parcialmente resolvido, utilizando dados de desambiguação obtidos da própria interface web da DBLP. Em alguns casos, a desambiguação foi manualmente resolvida. Isso ocorreu porque se tem um número restrito de pesquisadores sendo considerados (ver seção 4.5.1.3).

Outro desafio interessante é consolidar diferentes bibliotecas digitais como fontes, o que inclui questões de pesquisa, tais como: a conversão de diferentes formatos e/ou esquemas de dados; proveniência de dados; deduplicação de dados; desambiguação de nomes; entre outros (BILENKO; MOONEY, 2003; BORGES et al., 2011; CARVALHO et al., 2008; COHEN; RICHMAN, 2002; TEJADA; KNOBLOCK; MINTON, 2001). Outras fontes de dados, que contenham informações sobre publicações, como currículos e páginas pessoais dos pesquisadores, também podem ser utilizadas. Nesse caso, um importante desafio consiste em como suportar novos tipos de dados e algoritmos de processamento de consulta mais complexos. Especificamente, pode-se assumir que o conjunto de dados coletados sobre publicações e currículos irá conter dados não estruturados, semiestrutu-

[6]DBLP Computer Science Bibliography: http://www.informatik.uni-trier.de/~ley/db

rados (XML) e dados estruturados (relacionais). Esse tipo de conjunto de dados pode ser chamado de "dados híbridos" (MORO; LIM; CHANG, 2007). Tem-se consciência dos problemas citados acima, entretanto, eles não são o foco desta tese, e soluções apresentadas na literatura para esses problemas podem ser aplicadas.

4.5.1.2 Áreas de Pesquisa

A abordagem proposta considera as áreas de pesquisa nas quais os pesquisadores atuam (publicam) para determinar um indicador de "correlação". Os perfis dos pesquisadores, autores da rede de colaboração, foram construídos com base nas informações disponibilizadas na DBLP e em uma classificação das publicações dos autores feita pelo uso de uma ontologia de áreas de pesquisa, uma vez que a DBLP não disponibiliza essa informação. A ontologia para classificação dos artigos dos autores foi proposta por Loh et al. (2006). Essa ontologia utiliza uma classificação das áreas da Ciência da Computação similar à classificação da ACM (*Association for Computing Machinery*)[7]. Essa associa pesos a palavras-chave de acordo com suas relevâncias para representação de cada área de pesquisa. Para tanto, são utilizados vetores compostos por palavras-chave e seus pesos para representar a área de pesquisa (conceitos da ontologia) e o texto a ser classificado (artigos nesse caso). No vetor de uma área de pesquisa, cada peso corresponde à probabilidade de uma palavra-chave estar presente em um texto daquela área (LOH; WIVES; OLIVEIRA, 2000; LOH et al., 2010). No vetor de um artigo, cada peso é relativo à frequência de uma palavra-chave na informação do artigo. Então, o método de classificação consiste em avaliar a similaridade entre esses vetores. Nos experimentos apresentados nesta tese, os vetores dos artigos consideram apenas as palavras-chave presentes nos títulos das publicações.

Uma observação interessante é que a ontologia pode ser utilizada para determinar o "nível de detalhamento" da recomendação desejada em relação às áreas de pesquisas. Por exemplo, utilizar os níveis mais altos da classificação para recomendações mais "gerais" e os níveis mais baixos da classificação quando se deseja uma recomendação mais "específica". Essa pode ser uma flexibilidade bastante importante de ser explorada e permitida pelo uso desta abordagem. Por isso, pode ser bastante benéfico se trabalhar com uma ontologia (com uma classificação hierárquica) para classificar as áreas das publicações. A informação que a base contém sobre os artigos também pode determinar melhor o nível hierárquico mais adequado de ser explorado. No caso dos experimentos realizados no presente trabalho, lidou-se com níveis mais altos da classificação porque a DBLP disponibiliza apenas os títulos das publicações nos seus metadados, não permitindo, muitas vezes, uma classificação tão especificada das áreas das publicações.

4.5.1.3 Universo de pesquisadores

A fim de melhor avaliar a presente abordagem, foi limitado o universo de pesquisadores. Tal definição de limites foi efetuada para que os resultados obtidos pudessem ser mais detalhadamente analisados e verificados, pelo uso de um universo de pesquisadores conhecidos. Essa seleção também foi efetuada para testar o comportamento da abordagem em um subconjunto de pesquisadores (subrede). Além disso, uma subrede ou uma rede social incompleta é o caso mais provável de ocorrer, pela dificuldade se ter acesso a todos os pesquisadores e a todas as suas publicações. O conjunto de dados escolhido contém informações de publicações de 650 pesquisadores de programas de Pós-graduação

[7] ACM Computing Classification: http://www.acm.org/about/class/ccs98-html

Tabela 4.1: Programas de Pós-graduação em Ciência da Computação brasileiros selecionados para esta avaliação experimental.

Programa de Pós-graduação	**#pesquisadores**
COPPE/UFRJ	37
PUC/PR	20
PUC/RIO	24
PUC/RS	20
UFAM	16
UFBA	13
UFC	23
UFCG	27
UFF	58
UFMG	29
UFPE	58
UFPR	27
UFRGS	47
UFRJ	32
UFRN	25
UFSCAR	29
UNB	20
UNICAMP	43
USP	40
USP/SC	62
$\sum$	650

brasileiros da área da Ciência da Computação. A Tabela 4.1 apresenta tais programas de Pós-graduação em Ciência da Computação brasileiros (em ordem alfabética) e seus números de pesquisadores[8].

Dois intervalos de tempo foram escolhidos para serem utilizados nos experimentos: até 2007 e até 2010 (o intervalo de tempo é relacionado ao ano das publicações). Os intervalos de tempo foram escolhidos com o objetivo de analisar a evolução da rede de colaborações. Esse período é considerado consistente para o início de resultados de colaborações. Então, duas redes sociais formadas pelos pesquisadores dos programas de Pós-graduação em Ciência da Graduação brasileiros foram construídas, cada uma delas utilizando um dos dois intervalos de tempo descritos acima.

O número de atores de cada rede social é 650. Se todos os atores colaborassem entre si, o número de pares de vínculos relacionais (2 ligações entre dois autores em uma rede direcionada) na rede social seria 210.925. Na Tabela 4.2, são apresentadas algumas informações sobre o conjunto de dados considerado para a construção das Redes Sociais em cada intervalo de tempo. Nessa tabela, pode ser observado que as redes sociais correspondentes a ambos os conjuntos de dados (até 2007 e até 2010) contêm, respectivamente, 1.086 e 1.333 pares de vínculos relacionais. Um grande potencial para novas colaborações emergirem, nessa rede construída com dados até 2007, pode ser percebido (210.925

[8]As listas dos pesquisadores de cada programa de Pós-graduação foram obtidas das páginas dos programas das diferentes instituições, na mesma época da coleta dos dados (Agosto de 2010).

Tabela 4.2: Informação sobre o conjunto de dados dos programas de Pós-graduação em dois intervalos de tempo.

Informação	**Dados até 2007**	**Dados até 2010**
Total de publicações	10.735	14.563
Média de publicações por pesquisador	16,51	22,40
Número de pares de vínculos relacionais	1.086	1.333

- 1.086 = 209.839)[9].

Um estudo comparativo entre as duas redes sociais indicou que 247 novas colaborações (pares de vínculos relacionais) foram iniciadas no período considerado. Nos experimentos, estas novas colaborações foram consideradas relevantes se tivessem sido recomendadas em 2007, porque elas efetivamente ocorreram posteriormente. Existe uma evidência real de que estas colaborações são interessantes para os pesquisadores, porque publicação(ões) foi(foram) obtida(s) como resultados depois de 2007. Então, um bom sistema de recomendação deveria ser capaz de recomendar essas "novas" colaborações já em 2007. Tal cenário permite acuradamente avaliar a eficácia, em relação aos resultados desejados, de uma função de recomendação.

A Figura 4.10 mostra um gráfico resumido representando a Rede Social formada pelos pesquisadores dos programas de Pós-Graduação, utilizando o conjunto de dados até 2010. Por simplicidade, uma versão simplificada da rede é apresentada na qual os nós representam os pesquisadores e as arestas indicam a existência de coautorias entre pares de pesquisadores. Nesta versão simplificada, para melhor visualização, não estão representados a bi-direcionalidade e nem os pesos das ligações entre pesquisadores.

Em outro nível dos relacionamentos, a Figura 4.11 mostra um gráfico resumido representando a Rede Social agrupada por Programas de Pós-graduação. Tal gráfico foi construído, utilizando o conjunto de dados até 2010, e objetiva mostrar a cooperação entre os programas (cooperações interinstitucionais). No gráfico, instituições I_1 e I_2 são conectadas se pelo menos um dos pesquisadores da instituição I_1 tem uma colaboração com um pesquisador da instituição I_2. A espessura da linha representa a intensidade dos relacionamentos existentes. Esse gráfico é meramente ilustrativo, mostrando uma visão simplificada da Rede Social construída para conectar pesquisadores e mostra que várias análises diferentes podem ser realizadas, a partir da agregação e visualização de dados através técnicas específicas.

Finalmente, uma questão que pode ser levantada é sobre a competição entre pesquisadores. Neste trabalho, é assumido que a competição não é uma questão crítica. O conjunto de dados utilizado é composto por pesquisadores brasileiros pertencentes a programas de Pós-graduação. Nesse caso, uma competição individual não é crítica porque a posição dos pesquisadores dentro das universidades é relativamente estável. Além disso, em etologia[10], é um fato bem conhecido que um altruísmo individual é importante para melhorar a competitividade de um grupo como um todo. Dessa forma, é mais importante ter um grupo forte do que um indivíduo forte. Mais uma vez, esta tese focaliza nesses cenários nos quais a colaboração, em presença ou não de competição, é crucial.

[9]Tem-se consciência que esse valor está muito acima da realidade de possibilidades de colaboração, entretanto, mostra-se que a rede é esparsa e que selecionar pesquisadores para colaborar pode ser um desafio.

[10]Significado: Estudo dos costumes sociais humanos. (fonte: Dicionário Michaelis online: http://michaelis.uol.com.br/moderno/portugues/index.php?lingua=portugues-portugues&palavra=etologia)

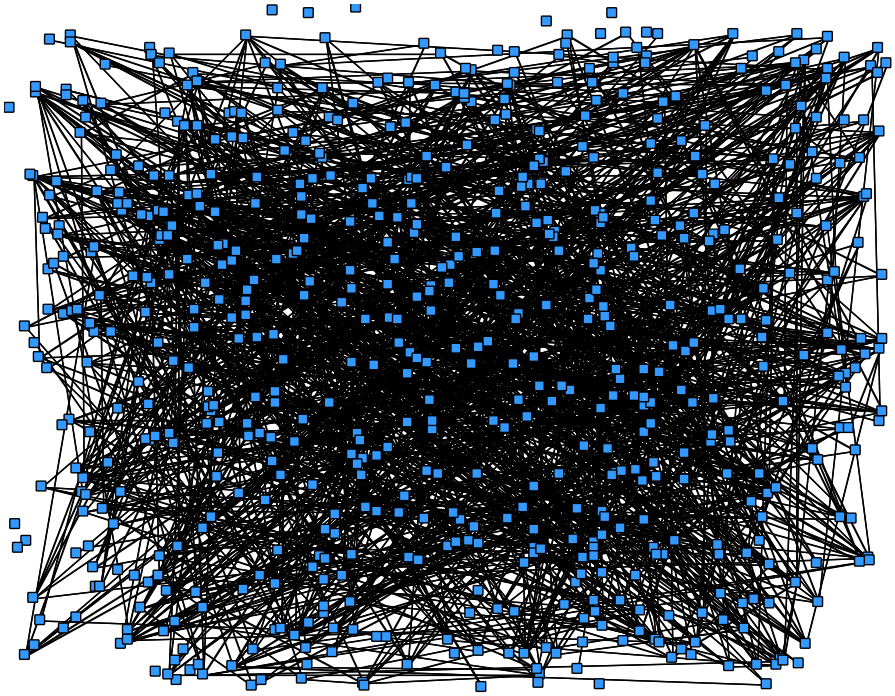


Figura 4.10: Rede Social formada pelos pesquisadores dos programas de Pós-graduação (conjunto de dados até 2010).

4.5.2 Experimentos Globais

Este conjunto de experimentos objetiva avaliar a abordagem de recomendação desta tese em relação a outras abordagens existentes utilizadas como *baseline*. Após apresentar as configurações (seção 4.5.1), nas seções seguintes, são apresentados e discutidos os *baselines* (seção 4.5.2.1) e os resultados (seção 4.5.2.2).

4.5.2.1 Abordagens utilizadas como Baseline

As outras abordagens escolhidas para serem utilizadas nesta avaliação experimental são a baseada em conteúdo (*content-based*) e o menor caminho (*shortest path*), bem como suas combinações. Essas duas abordagens foram escolhidas, porque elas são utilizadas em outras propostas e avaliações do estado da arte tais como (CHEN et al., 2009; LIBEN-NOWELL; KLEINBERG, 2007) e cobrem certas facetas da solução do problema proposta pela abordagem da presente tese.

A abordagem adotada, baseada em conteúdo (*content-based*), considera o modelo espaço vetorial (*vector space model*-VSM) (SALTON; BUCKLEY, 1988). Nesse modelo os perfis dos usuários são representados como vetores de termos de indexação. O modelo espaço vetorial utiliza um espaço t-dimensional para representar os termos, no qual t corresponde ao número de termos distintos. Para cada vetor do perfil do usuário, os pesos representam as coordenadas do vetor na dimensão correspondente. Os termos que compõem o perfil do usuário são obtidos dos títulos das publicações do pesquisador. No processo de construção do perfil do usuário, as *stopwords* são removidas (uma lista de termos muito comuns ou gerais que não são usados no processo de recuperação de informação, por exemplo, preposições, conjunções e artigos). Além disso, cada termo tem um peso associado. Esse peso indica a importância do termo para a representação do perfil do usuário e é calculado de acordo com a Equação 4.13.

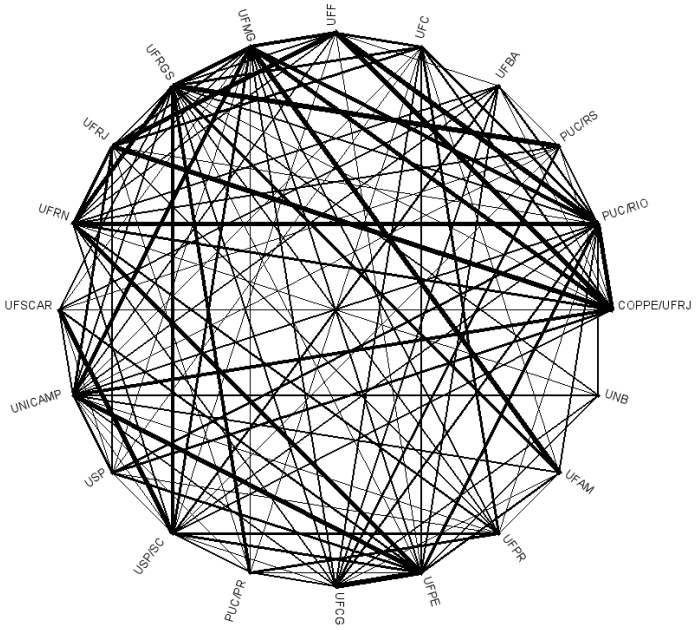


Figura 4.11: Rede Social agrupada por programas de Pós-graduação (conjunto de dados até 2010).

$$w_{t_i} = \frac{tf_i}{\sum_{j=1}^{t} tf_j} \tag{4.13}$$

Nessa equação, o peso de cada termo i é determinado pela frequência do termo i (tf_i) dividido pela frequência total dos termos no perfil do usuário sendo considerado ($\sum_{j=1}^{t} tf_j$). Uma vez que os vetores representando os perfis dos usuários são construídos, é possível calcular os escores de similaridade entre eles, através da equação do cosseno. Essa equação provê uma saída ranqueada com base na ordem decrescente dos valores de similaridade (SALTON; BUCKLEY, 1988). Então, essa saída representa o resultado de uma função de recomendação genérica baseada em conteúdo.

O menor caminho (*shortest path*) é um método amplamente utilizado para predição de ligações (LIBEN-NOWELL; KLEINBERG, 2007), bem como é empregado no contexto da recomendação de amigos (CHEN et al., 2009). Nesse caso, uma rede social simplificada *SN* das relações de coautoria a é um par: $SN_a = (N, E)$ onde N e E são o conjunto de nós (Nodes) e arestas (Edges), respectivamente. Cada aresta $e \in E$ é não-direcionada, conectando dois autores (a_i e a_j), sendo a_i e a_j autores diferentes, e w o peso afetado pela associação. Se um relacionamento entre dois autores existe, o peso é o valor numérico 1; caso contrário, o peso é ∞. Essa métrica segue a noção de mundos pequenos, no qual indivíduos são relacionados através de curtas cadeias de relacionamentos (poucos graus de separação). A saída resultante é uma recomendação baseada na métrica do menor caminho.

As abordagens baseada em conteúdo (*content-based*) e menor caminho (*shortest path*) também foram combinadas para serem utilizadas como *baselines* otimizados. Especificamente, foi utilizada a função de intersecção de conjuntos e uma das métricas foi escolhida para ordenar os resultados de recomendação. Então, um total de quatro abordagens foram utilizadas como *baseline* na avaliação comparativa: baseada em conteúdo (*content-based*)

referenciada por "CB"; menor caminho (*shortest path*) referenciada por "SP"; intersecção entre baseada em conteúdo (*content-based*) e menor caminho (*shortest path*) ordenada pelo valor da métrica baseada em conteúdo (*content-based*) referenciada por "CB ∩ SP (ordered by CB)"; e intersecção entre baseada em conteúdo (*content-based*) e menor caminho (*shortest path*) ordenada pelo valor da métrica menor caminho (*shortest path*) referenciada por "CB ∩ SP (ordered by SP)".

4.5.2.2 Resultados da Avaliação Comparativa

A fim de desenvolver uma avaliação comparativa, foram utilizadas todas as recomendações possíveis para todos os membros dos programas de Pós-graduação em Ciência da Computação brasileiros (programas selecionados para esta avaliação experimental) geradas pelas abordagens, usando o conjunto de dados no primeiro intervalo de tempo (publicações até 2007). Também foram comparadas as modificações encontradas entre as redes nos dois intervalos de tempo considerados (até 2007 e até 2010) e as recomendações geradas. A evolução real ocorrida quando comparadas as redes com dados até 2007 e até 2010 foi considerada o resultado correto (recomendações relevantes que deveriam ter sido efetuadas).

Esta avaliação experimental focaliza na recomendação de novas colaborações, porque é possível mensurar pela comparação da rede no passado com a situação corrente (como feito considerando os anos de 2007 e 2010). O outro caso (intensificação de colaborações) é praticamente impossível de acuradamente verificar, como em ciências sociais aplicadas, é muito difícil criar um grupo de referência idêntico ao grupo experimental que sofreu uma intervenção, tal que se possa, explicitamente, medir as melhorias, sofridas pelo grupo, causadas pelo uso da abordagem de recomendação. O modo adequado de fazer isso é previamente recomendar as intensificações a um grupo e verificar os impactos sofridos pelo mesmo em um momento posterior. Note que isso deve ser também comparado contra outro grupo que não sofreu qualquer intervenção experimental. O ideal seria que o grupo que recebeu as recomendações obtivesse melhorias/ganhos em relação ao que não as recebeu.

Os resultados dos experimentos foram realizados por meio de estratégias de avaliação padrão na área de Recuperação de Informações. Nesse contexto, as seguintes métricas foram utilizadas: precisão (*precision*), revocação (*recall*), média das precisões médias (*mean average precision - MAP*), precisão até 10 (*precision at 10 - Pr@10*) e Teste T (*Student's t-test*) (BAEZA-YATES; RIBEIRO-NETO, 2011; MANNING; RAGHAVAN; SCHÜTZE, 2008)[11]. Além desses, um método referenciado aqui como *DifferenceScores*, sugere que a diferença nos escores das estratégias de avaliação entre dois sistemas pode ser considerada notável quando for maior do que 5% e pode ser considerada material quando for superior a 10%. De acordo com Buckley e Voorhees (2000), esse método é baseado em uma noção razoável de diferença e é um método padrão utilizado na avaliação de sistemas de recuperação de informação, sendo usado em campanhas de avaliação como CLEF (*Cross Language Evaluation Forum*)[12] e TREC (*Text REtrieval Conference*)[13].

Análise dos parâmetros.

[11]Como estes experimentos irão mostrar, o valor de precisão é usado para indicar se uma rede de colaboração pode se beneficiar pelo uso de uma função de recomendação (não se uma função de recomendação está correta). Dessa forma, a métrica de F1 não é considerada, uma vez que essa é uma média harmônica entre revocação e precisão.

[12]CLEF: http://clef-campaign.org

[13]TREC: http://trec.nist.gov

Tabela 4.3: Resultados de revocação (Recall), média das precisões médias (MAP) e precisão até 10 (Pr@10) para o conjunto de dados dos Programas de Pós-graduação em Ciência da Computação brasileiros.

	Método	**Revocação**	**MAP**	**Pr@10**
Parte A	Cr	75,71%	1,17%	0,43%
	Sc	74,89%	6,35%	2,14%
Parte B	$Cr \cap Sc$ (ordered by Cr)	65,99%	1,32%	0,48%
	$Cr \cap Sc$ (ordered by Sc)	65,99%	6,12%	1,91%
	Cr_Sc (w_{Cr}=1, w_{Sc}=150)	**84,61%**	**6,72%**	**2,17%**

O primeiro caso considerado pela abordagem de recomendação desta tese é quando dois pesquisadores não têm artigos em coautoria, ou seja, não precisa considerar a métrica de cooperação diretamente. Nesta seção, os resultados obtidos pelas métricas Cr e Sc são analisados a fim de gerar recomendações ranqueadas. Como efeito colateral, este estudo foi conduzido, objetivando ajudar na associação dos valores dos pesos w_{Cr} e w_{Sc} utilizados na abordagem Cr_Sc (no restante dos experimentos desta seção). A Tabela 4.3 sumariza os resultados de revocação (recall), média das precisões médias (MAP) e precisão até 10 (Pr@10), para as métricas individuais e combinadas (Cr e Sc utilizadas separadamente e combinadas, utilizando a intersecção - ordenada por uma das abordagens - e a união de seus resultados pela abordagem Cr_Sc).

Como mostrado pela Equação 4.6, o escore final de recomendação depende dos pesos definidos para correlação e proximidade social. De acordo com a Tabela 4.3 Parte A, dentre Cr e Sc, o último apresentou o maior valor de precisão. Então, o peso w_{Sc} deve ser maior do que o de w_{Cr}. O próximo experimento objetiva ajudar na associação de tais pesos. Com uma determinação adequada, a abordagem deste trabalho pode obter maiores valores de precisão.

Dado que Sc se mostrou mais importante, a Tabela 4.4 apresenta os valores de precisão MAP e Pr@10 obtidos variando os valores de w_{Sc} para um valor fixo de $w_{Cr} = 1$. Especificamente, a Tabela 4.4 mostra que MAP atinge seu valor máximo e estabiliza em torno de $w_{Sc} = 100$. A tabela também mostra que o valor de Pr@10 atinge seu valor máximo e estabiliza em torno de $w_{Sc} = 150$. Então, o restante da avaliação comparativa tem o valor padrão (*default*) de 150 para o peso w_{Sc}.

É importante notar que a Tabela 4.3 Parte A também provê evidências da importância da *proximidade social* para o estabelecimento de novas colaborações. Em outras palavras, para atingir resultados mais acurados não é suficiente considerar apenas a correlação. No entanto, tem-se consciência de que os resultados foram avaliados, dependendo da atual evolução da rede, considerando o ano de 2010. Esse resultado também mostra que a *proximidade social* é adequadamente estimada pela métrica Sc.

Os valores de MAP e Pr@10 obtidos por Cr_Sc utilizando $w_{Cr} = 1$ e $w_{Sc} = 150$ são apresentados na Tabela 4.3 Parte B. Pode-se concluir que, para esse caso, a abordagem Cr_Sc ajudou a melhorar os valores de revocação quando comparados com os outros resultados. Além disso, podem ser obtidos valores de pesos que não diminuem a precisão e até mesmo obtêm uma ligeira melhora quando comparados aos valores isolados de Cr e Sc. Esse é um estudo preliminar da associação dos pesos w_{Cr} e w_{Sc}. A determinação dos pesos utilizados na abordagem Cr_Sc depende do tipo de recomendação que o usuário espera. Dessa forma, a associação de diferentes pesos para diferentes usuários pode

Tabela 4.4: Efeito da variação do peso w_{Sc} nos resultados da ordenação das recomendações analisado por média das precisões médias e precisão até 10.

w_{Sc}	**MAP**	**Pr@10**
0	1,17%	0,43%
1	2,85%	0,98%
25	6,03%	1,98%
50	6,46%	2,05%
75	6,64%	2,03%
100	6,79%	2,11%
125	6,73%	2,15%
150	6,72%	2,17%
175	6,71%	2,17%
...	...	...

Tabela 4.5: Resultados de revocação e precisão para o conjunto de dados dos programas de Pós-graduação em Ciência da Computação brasileiros.

Método	**Revocação**	**Precisão**
Cr_Sc	**84,61%**	**0,1380%**
Content-based (CB)	64,37%	0,1449%
Shortest path (SP)	74,89%	0,1544%
CB ∩ SP	54,25%	0,1721%

ser mais adequada. Alguns usuários podem preferir recomendações de pesquisadores em uma maior proximidade social com eles (peso w_{Sc} maior) enquanto outros podem preferir recomendações de pesquisadores que têm uma maior similaridade no perfil de atuação, considerando as áreas de pesquisa (peso w_{Cr} maior). Dessa forma, destaca-se a flexibilidade permitida pela abordagem desta tese. Em trabalhos futuros, pode ser analisada a escolha desses pesos, utilizando *feedback* do usuário.

Os resultados da Tabela 4.3 Parte B indicam que o maior valor de revocação foi obtido pela abordagem Cr_Sc. As diferenças entre a revocação obtida por Cr_Sc e por todas as outras abordagens foi maior do que 9% (em todos os casos), que pode ser considerada uma diferença estatisticamente significativa. Esses resultados mostram evidências da importância das métricas de *proximidade social* (Sc) e *correlação* (Cr) na representação de facetas que devem ser levadas em conta para o estabelecimento de novas colaborações no contexto acadêmico.

Avaliação de Revocação e Precisão.

Anteriormente (nessa seção), foi apresentada uma análise dos parâmetros utilizados pela função de recomendação, ou seja, os pesos para correlação e proximidade social. A seguir, aqueles parâmetros são usados para avaliar a função de recomendação desta tese *versus* os *baselines*. A Tabela 4.5 apresenta os resultados de revocação (*recall*) e precisão (*precision*).

A avaliação considera todas as recomendações (pares de autores) cujos valores de escore são maiores do que 0. Pode-se observar que os valores de revocação obtidos pela abordagem desta tese (Cr_Sc) foram maiores do que os das outras abordagens considera-

das na avaliação comparativa. As diferenças entre a abordagem Cr_Sc e todas as outras foram maiores do que 9% (em todos os casos), e estas diferenças podem ser consideradas estatisticamente significativas (preliminarmente provado pelo método *DifferenceScores*).

Tais resultados reforçam a hipótese de que a abordagem de recomendação desta tese: levando em consideração a similaridade entre os perfis de atuação dos usuários por áreas de pesquisa e a consideração da *proximidade social*, pode conduzir a melhorias significativas na revocação das recomendações. Na verdade, pesquisadores com perfis similares podem ter mais chances de virem a cooperar.

A precisão de todas as abordagens apresentou baixos valores. Isso provavelmente ocorreu porque as recomendações não foram avaliadas pelos usuários. Em vez disso, apenas as novas colaborações, que emergiram no período avaliado (2008-2010) e resultaram na publicação de artigo, é que foram consideradas relevantes. Essa avaliação de precisão objetiva mostrar se existem evidências significativas da importância de sistemas de recomendação no contexto de redes sociais acadêmicas, considerando duas hipóteses como segue.

Primeira, se o valor de precisão for alto, considerando somente as novas colaborações que emergiram naturalmente na rede social acadêmica, então a rede já está bem. Em outras palavras, já está se comportando de maneira a otimizar seu potencial de colaboração. Nesse caso, um sistema de recomendação pode não ser necessário nem útil. Segunda, se o valor de precisão for baixo, um sistema de recomendação pode ser fundamental para aumentar o potencial de colaboração. Nesse outro caso, um sistema de recomendação será muito mais vantajoso, útil e importante.

Os resultados experimentais mostraram baixos valores de precisão para todas as abordagens, conforme apresentado na Tabela 4.5. Esse resultado é uma evidência da importância e utilidade do desenvolvimento e aplicação de sistemas de recomendação no contexto acadêmico. Além disso, os resultados mostram que a abordagem desta tese pode encontrar um maior número de recomendações se não existir a escolha de um número máximo de recomendações ou o estabelecimento de um limiar (*threshold*) para o valor do escore de recomendação.

Curvas de Revocação-Precisão e testes estatísticos.

Esta segunda avaliação apresenta os pontos de precisão interpolada (normalizada) das abordagens analisadas, utilizando 11 pontos padrão de revocação, que é uma estratégia padrão de avaliação de sistemas de recuperação de informações (BAEZA-YATES; RIBEIRO-NETO, 2011; MANNING; RAGHAVAN; SCHÜTZE, 2008). Com estas curvas de revocação interpolada e normalizada, é possível entender o comportamento geral de um sistema.

A Figura 4.12 mostra curvas de revocação-precisão para cada uma das abordagens. Como comentado anteriormente, as recomendações são consideradas relevantes se elas representam uma colaboração que surgiu no período considerado. Estas novas colaborações podem representar evidências significativas do conjunto de recomendações ótimo. Então, uma boa abordagem de recomendação deve ranquear estas novas colaborações mais no topo em relação às outras recomendações.

Pode-se observar que os valores de precisão obtidos pela abordagem desta tese (Cr_Sc) foram em média maiores que os resultados das outras abordagens para todos os níveis de revocação. Os valores da média das precisões médias (MAP) e da precisão até 10 (Pr@10) obtidos pelas abordagens consideradas estão presentes na Tabela 4.6.

Uma observação importante é que as abordagens SP e CB $\cap$ SP (ordered by SP) podem gerar uma série de escores empatados, porque todos os pesquisadores "separados"

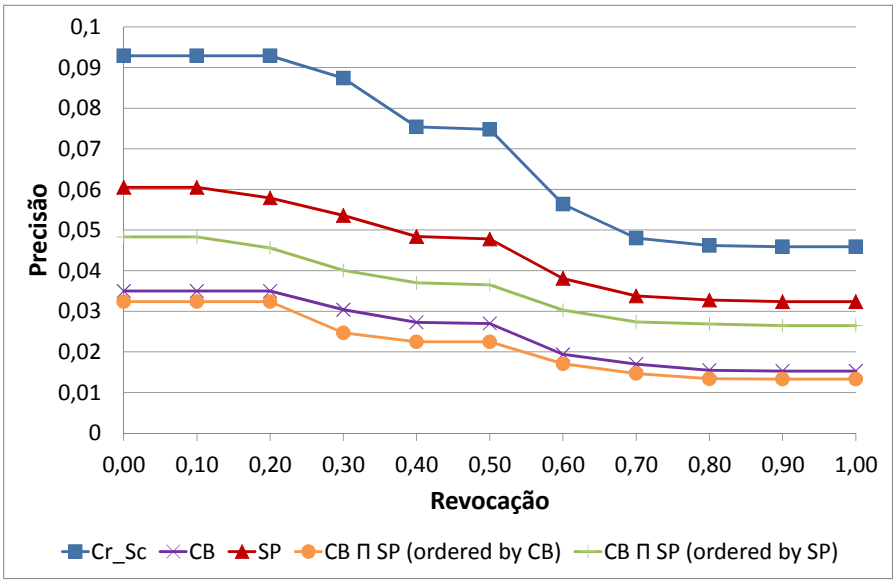


Figura 4.12: Curvas de Revocação-Precisão das diferentes abordagens de recomendação para os programas de Pós-graduação em Ciência da Computação brasileiros.

Tabela 4.6: Resultados da Média das Precisões Médias (MAP) e Precisão até 10 (Pr@10) para o conjunto de dados dos programas de Pós-graduação em Ciência da Computação brasileiros. Valores entre colchetes indicam resultados utilizando métricas específicas para lidar com empates.

Método	**MAP**	**Pr@10**
Cr_Sc (w_{Cr}=1, w_{Sc}=150)	**6,72%**	**2,17%**
Content-based (CB)	2,38%	0,83%
Shortest path (SP)	4,32% [4,37%]	1,52% [1,79%]
CB ∩ SP (ordered by CB)	2,08%	0,68%
CB ∩ SP (ordered by SP)	3,41% [3,58%]	1,22% [1,41%]

pelo mesmo número de pesquisadores em relação ao usuário alvo terão o mesmo valor de escore de recomendação. Essas abordagens efetuam apenas uma ordenação parcial nos resultados porque existem múltiplas ordenações possíveis no ranking final de recomendação, cada uma efetuada de forma diferente. Os resultados apresentados na Figura 4.12 foram gerados, utilizando o id (*identification*) dos pesquisadores na base de dados como critério de desempate para determinar a ordem das recomendações que apresentam escores empatados. Na Tabela 4.6, os valores das abordagens SP e CB ∩ SP (ordered by SP) utilizando os id's como critério de desempate são apresentados.

Além disso, existem métricas desenvolvidas especificamente para lidar com empates, chamadas *Tie-Aware*. Métricas *Tie-Aware* devem calcular a média dos valores de performance através de todos os possíveis resultados de ordenação. Métricas *Tie-Aware* para determinar MAP e Pr@10 propostas por McSherry e Najork (2008) são também utilizadas na avaliação das duas abordagens citadas acima (SP e CB ∩ SP (ordered by SP)). Na Tabela 4.6, os valores de MAP e Pr@10 calculados utilizando métricas *Tie-Aware* são apresentados entre colchetes.

A abordagem desta tese (Cr_Sc) foi comparada com os quatro *baselines*. A fim de

avaliar se as melhorias apresentas na Tabela 4.6 são estatisticamente significativas, o teste estatístico T foi utilizado. De acordo com Hull (1993), o teste T funciona bem, até mesmo com distribuições não perfeitamente normais. Um limiar padrão de significância estatística α de 0,05 foi utilizado. Quando o valor de p calculado for menor que α, existe uma diferença estatisticamente significativa entre as abordagens comparadas. Os resultados do teste estatístico mostraram que a abordagem Cr_Sc alcançou melhorias estatisticamente significativas quando comparado com todos os *baselines*. Dessa forma, o teste estatístico T comprovou que a abordagem Cr_Sc obteve um desempenho significativamente superior em relação a todas as quatro abordagens utilizadas como *baseline*.

4.5.3 Experimentos sobre os Refinamentos com Aspectos Temporais

Estes experimentos objetivam considerar diferentes métricas para ponderação dos vínculos relacionais de forma a avaliar a influência dos aspectos temporais na recomendação de colaborações em redes sociais acadêmicas (conforme seção 4.3). O conjunto de dados para a construção das redes estudadas, nesta seção, foi descrito previamente na seção 4.5.1 e os resultados da avaliação estão descritos a seguir.

Nestes experimentos, é adotada uma função de recomendação simplificada para recomendação de novas colaborações (apresentada previamente na seção 4.3). Nesse caso, foi adotada a função de escore (para recomendação) como o menor caminho porque a ideia é de que esta é muito dependente dos pesos que são dados às arestas, sendo uma boa forma de medir melhor o impacto da modificação desses pesos. Tal função foi adotada, porque aqui a intenção é avaliar o impacto da consideração de diferentes métricas para ponderar os pesos da rede social acadêmica (refinadas para considerar aspectos temporais) no resultado das recomendações. Dessa forma, para avaliar o real impacto sofrido pela abordagem de recomendação em consequência desses refinamentos, essa função precisa utilizar apenas o indicador de proximidade social.

A fim de desenvolver uma avaliação comparativa, foram utilizadas todas as possíveis recomendações definidas pelas abordagens, utilizando o conjunto de dados no primeiro intervalo de tempo (publicações até 2007). Também foram comparadas as novas coautorias do segundo conjunto de dados (até 2010) e as recomendações geradas pelas abordagens. Como grande verdade, da mesma forma que na avaliação anterior, foi considerada a evolução real de 2007 até 2010, ou seja, as atuais novas conexões que aconteceram através do tempo foram consideradas como o resultado correto.

Os resultados dos experimentos são avaliados por estratégias de avaliação padrão em recuperação de informações. Especificamente, as seguintes métricas foram utilizadas: precisão média (*average precision* - Avg-Prec), média das precisões médias (*mean average precision* - MAP), precisão até R (*r-precision* - R-prec), revocação (*recall*) e teste T (*Student's t-test*) (BAEZA-YATES; RIBEIRO-NETO, 2011). Para avaliação das métricas, cada usuário foi considerado como uma consulta e cada recomendação de colaboração para o usuário como o resultado daquela consulta. Uma breve explicação sobre as métricas de avaliação associadas ao domínio é apresentada abaixo.

- **Precisão** é a fração de recomendações recuperadas que são relevantes para o usuário.
- **Precisão média** é a média das precisões computadas para cada ponto de recomendações relevantes na sequência de ranqueamento.
- **Média das precisões médias** é calculada para um conjunto de usuários como a

média dos escores de precisão média para cada usuário.

- **Precisão até R** para um conjunto de usuários é a média das precisões calculadas até a posição R do ranking de resultados (onde R representa o número de recomendações relevantes para cada usuário).

- **Revocação** é a fração das recomendações que são relevantes para o usuário que foram recuperadas com sucesso.

- **Teste T** é um teste estatístico empregado para verificar a existência de diferenças estatisticamente significativas.

A Tabela 4.7 lista as métricas, os valores dos parâmetros e os resultados de avaliação. As métricas de estabelecimento de pesos avaliadas nestes experimentos são apresentadas a seguir.

1. Na primeira configuração, nenhuma métrica para estabelecimento de pesos é usada e os vínculos relacionais não são ponderados.

2. A segunda métrica é uma variante assimétrica do coeficiente de Jaccard (apresentada na Equação 4.9).

3. A terceira métrica utiliza somente o fator temporal (conforme Equação 4.10), utilizando os parâmetros especificados na Tabela 4.7.

4. A quarta métrica é a Tg (são apresentados os resultados para duas diferentes combinações de parâmetros) que considera o uso do fator temporal globalmente, utilizando os anos de publicação de todos os artigos.

5. A quinta métrica é a Tr (são apresentados os resultados para três diferentes combinações de parâmetros) que considera o fator temporal aplicado utilizando somente o ano mais recente dentre todos os anos de publicações em coautoria entre dois pesquisadores.

Como a Tabela 4.7 mostra, todas as configurações (métricas com o uso dos parâmetros indicados) proveram baixos valores de MAP e R-prec. Isso provavelmente ocorreu porque as recomendações não foram avaliadas pelos usuários e somente as novas colaborações que emergiram no período avaliado (2008-2010) é que foram consideradas relevantes. Os resultados de MAP da melhor configuração não considerando aspectos temporais (#2) e todas as outras utilizando aspectos temporais mostram que não existem diferenças estatisticamente significativas entre elas, quando utilizado um teste T com nível de significância de 0,05. Isso provavelmente ocorreu porque a mesma função de escore (menor caminho) está sendo utilizada em todos os casos. Então, as mesmas recomendações relevantes estão sendo retornadas pela maioria das configurações, e a diferença é somente na ordenação dos resultados. A única exceção é a configuração numerada como #5.3 (Tr com $t_{min} = 0$) na qual existe uma redução nos relacionamentos representados na Rede Social, porque as relações dos pesquisadores com colaborações mais antigas que 10 anos (valor de w) não serão modeladas (valor de $t_{min} = 0$).

Neste estudo de caso, a redução no número de relacionamentos na Rede Social pelo uso da configuração #5.3 foi de aproximadamente 4,24%. Então, o número de relacionamentos da Rede Social considerado pelo método de escore para gerar as recomendações

Tabela 4.7: Resultados de Média das precisões médias, Precisão até R e Revocação para diferentes configurações (métricas de estabelecimento de pesos e seus respectivos parâmetros atribuídos).

# Configuração	Métrica	Parâmetros temporais	MAP	R-prec	Revocação
1	Sem pesos	Não aplicável	4,34%	2,51%	74,09%
2	Quantidade de publicações	Não aplicável	6,36%	4,65%	74,09%
3	Somente fator temporal	$[y_r = 2007, w = 10, t_{min} = 0,01]$	4,08%	2,65%	74,09%
4.1	Tg	$[y_r = 2007, w = 10, t_{min} = 0,01]$	6,59%	4,71%	74,09%
4.2	Tg	$[y_r = 2007, w = 20, t_{min} = 0,01]$	6,38%	4,58%	74,09%
5.1	Tr	$[y_r = 2007, w = 10, t_{min} = 0,01]$	6,72%	5,03%	74,09%
5.2	Tr	$[y_r = 2007, w = 20, t_{min} = 0,01]$	6,55%	4,81%	74,09%
5.3	Tr	$[y_r = 2007, w = 10, t_{min} = 0,00]$	6,56%	5,07%	72,29%

também foi reduzido. Apesar dessa redução, os resultados de MAP da configuração #5.3 foram ligeiramente superiores quando comparados com o melhor resultado sem considerar os aspectos temporais (configuração #2). Além disso, esses resultados mostram que aspectos temporais podem, na verdade, ser usados para reduzir o espaço de busca e ainda assim obter resultados comparáveis àqueles sem a redução.

Em termos de revocação (veja Tabela 4.7), a configuração #5.3 reduziu o valor de revocação da abordagem de menor caminho de 74,09% para 72,29% com melhorias em MAP e R-prec quando comparados com as configurações não utilizando aspectos temporais. Neste experimento, a Rede Social considerada já sofreu uma seleção nos seus relacionamentos (os atores são um conjunto de professores de programas de Pós-graduação). A existência de colaborações antigas ou descontinuadas entre esses professores pode não ocorrer tão frequentemente quanto em outros cenários. Por exemplo, colaborações descontinuadas ocorrem mais frequentemente entre professores e seus estudantes, uma vez que alguns estudantes se formam e podem não seguir na área acadêmica. Nesse caso, uma redução nos relacionamentos, utilizando os aspectos temporais da Rede Social, pode ser ainda mais vantajosa e conduzir a melhorias ainda mais significativas nos resultados das recomendações.

Uma vez que os aspectos temporais não proveram muitas melhorias em termos de MAP, foram analisados os resultados de Avg-prec para cada pesquisador. Uma análise, usuário por usuário, verificou os casos em que considerar aspectos temporais melhora ou piora os resultados de recomendação em termos de Avg-Prec (somente considerando os casos nos quais pelo menos um dos métodos retornou resultados de recomendação relevante). A Tabela 4.8 sumariza os resultados. Ocorreram 91 melhoras, 58 pioras e 49 resultados sem alteração em termos dos valores de Avg-Prec das recomendações dos usuários. Então, para a configuração #5.1, as melhoras por usuário ocorreram em 45,96% dos casos e as pioras em somente 24,75% destes. Esses resultados são ainda melhores quando o aspecto temporal com redução da Rede Social é considerado (configuração #5.3). Para este, ocorreram 115 melhoras, 50 pioras e 33 resultados sem alteração em termos de Avg-Prec das recomendações dos usuários. Então, para a configuração #5.3, as melhoras por usuário ocorrem em 58,08% (maioria, mais da metade) dos casos e as pioras somente em 16,67% destes. Esses resultados apontam indícios das melhorias na ordenação dos resultados das recomendações pela consideração de aspectos temporais.

A Tabela 4.9 apresenta as cinco maiores melhoras e pioras da configuração #5.1 (Tr) em comparação com a configuração #2 (pesos somente considerando aspectos de quantidade de publicações). Nessa tabela, as cinco maiores melhoras são representadas por valores de diferenças positivas, enquanto que as cinco maiores pioras são representadas

Tabela 4.8: Comparativo considerando os usuários para os quais pelo menos uma das abordagens retornou alguma recomendação relevante.

	Configuração #5.1		**Configuração #5.3**	
	# usuários	**percentual**	**# usuários**	**percentual**
Melhora	91	45,96%	115	58,08%
Piora	58	29,29%	50	25,25%
Sem alteração	49	24,75%	33	16,67%

Tabela 4.9: 5 primeiros usuários mais beneficiados e prejudicados com o uso de aspectos temporais.

# pesquisador	**Configuração #2**	**Configuração #5.1**	**Diferença**
	Avg-Prec	**Avg-Prec**	
2	50,00%	100%	+50,00%
21	23,81%	66,67%	+42,86%
166	16,67%	50,00%	+33,33%
624	18,33%	41,67%	+23,34%
634	12,50%	33,33%	+20,83%
447	50,00%	33,33%	-16,67%
516	35,84%	19,18%	-16,66%
37	28,57%	12,04%	-16,53%
591	33,33%	25,00%	-8,33%
605	14,29%	7,69%	-6,60%

por diferenças negativas. Pode ser observado que o percentual de melhora (em valores absolutos) foi maior que os valores de percentual de pioras.

Nesse caso, as recomendações para o usuário numerado como 2 obtiveram o maior percentual de melhoria. Por exemplo, a segunda maior melhora ocorreu com o usuário 21. Para este usuário, a configuração #2 retornou as recomendações relevantes nas posições 3 e 9 do ranking, enquanto a configuração #5.1 retornou as recomendações relevantes nas posições 1 e 6 do ranking. Então, o uso de aspectos temporais melhorou todas as posições das recomendações nesse caso. Para ilustrar, a Rede Social parcial da Figura 4.7(a) representa as conexões da rede usadas para gerar as três primeiras recomendações da configuração #2 para o usuário alvo 21 (pesquisadores 25, 202 e 143 nessa ordem). Entretanto, pelo uso da configuração #5.1, a ordem dessas recomendações foi modificada (pesquisador 143 primeiro, 202 como segundo e 25 somente na posição vinte e cinco). Dentre esses três pesquisadores, o único considerado como recomendação relevante para o usuário alvo 21 é o usuário 143. Este pesquisador, pelo uso de aspectos temporais, foi corretamente ordenado como o primeiro. Em um segundo exemplo, a maior piora ocorre com o usuário 516. Para este usuário, a configuração #2 retornou recomendações relevantes nas posições 1, 32 e 239 do ranking; enquanto que a configuração #5.1 retornou as recomendações relevantes nas posições 2, 32 e 235 do ranking. Para este usuário, o ganho em Avg-Prec da configuração #2 ocorreu somente porque a primeira recomendação relevante é apresentada na posição 1; para a ordenação das duas outras recomendações, a configuração #5.1 tem uma performance igual ou superior. Esses resultados são muito

interessantes e mostram que considerar aspectos temporais pode levar a melhorias na ordenação de recomendações.

Neste capítulo, foi apresentada uma abordagem inovadora para recomendar colaborações científicas no contexto de redes sociais acadêmicas. Foram explorados três indicadores principais nomeados cooperação, correlação e proximidade social, para definir uma função de recomendação com diferentes recomendações de colaborações, novas e a serem intensificadas. Além disso, foram definidas diferentes métricas para tentar estimar esses indicadores. Foram explorados refinamentos na métrica de cooperação que visam a considerar aspectos temporais das relações da rede social acadêmica utilizada como base para gerar recomendações. Além disso, foi apresentada uma série de experimentos efetuados para validar e avaliar os resultados da abordagem proposta. No capítulo seguinte, são retomados os objetivos e destacadas as principais contribuições desta tese como um todo.

5 CONCLUSÕES

Neste capítulo, são apresentadas as considerações finais, destacando-se a sumarização das contribuições e resultados obtidos por esta tese (seção 5.1). Dentre as contribuições e resultados, são apresentadas as publicações resultantes dos trabalhos realizados no decorrer do doutorado. Por fim, são discutidos alguns trabalhos futuros identificados (seção 5.2).

5.1 Contribuições

Nesta tese, foram apresentadas propostas para avaliação e recomendação de colaborações científicas no contexto de Redes Sociais. Foi discutido que esta tese focaliza em cenários onde a colaboração é vital. Sem a perda de generalidade, destaca-se que o presente trabalho é flexível e generalizável para diferentes áreas. Mais ainda, é muito importante notar que ele pode ser aplicado para um cenário industrial, por exemplo. Nesse caso, ao invés de somente considerar os artigos produzidos por um grupo, a abordagem apresentada nesta tese pode ser facilmente adaptada para lidar com patentes ou outros tipos de informação, desde que exista um modo de estabelecer cooperações prévias entre indivíduos e suas especialidades por áreas.

Inicialmente, foi apresentada uma ampla revisão bibliográfica, incluindo os assuntos permeados por esta tese e comparativos entre os principais trabalhos relacionados e as novas proposições.

Foi proposta uma nova função para avaliação de qualidade de grupos de pesquisadores. Foi considerada a hipótese de que grupos de pesquisadores que são internamente colaborativos têm mais chance de alcançar sucesso e excelência em pesquisa do que grupos sem atividade social. Mais ainda, foram propostas métricas para, adequadamente, quantificar indicadores de qualidade, considerando essa hipótese para geração de rankings.

Foram desenvolvidos experimentos, usando um conjunto de dados reais dos programas de Pós-graduação brasileiros. A análise mostrou que pesquisadores de programas de melhor qualidade têm a tendência a apresentar um maior comportamento colaborativo. Foi estabelecida uma análise comparativa, utilizando uma avaliação oficial desenvolvida pela CAPES como *baseline*. Os resultados mostraram evidências de que um importante indicador de qualidade para programas de Pós-graduação é a análise da colaboração interna. Além disso, as novas métricas propostas foram apropriadas para quantificar indicadores de qualidade com propósitos de geração de rankings. É enfatizado, e estudado mais aprofundadamente, que a proposta de uso da avaliação do maior autovalor obteve uma excelente correlação com o ranking de classificação da CAPES. São destacados os resultados obtidos pelas métricas ineficiência social e maior autovalor (uso do Coeficiente de Spearman para comparação entre rankings e teste de significância estatística). A

ineficiência social aliou bons resultados com a simplicidade de cálculo. Já pela análise do maior autovalor da matriz de adjacência de um grafo (representando a rede de colaboração de um programa de Pós-graduação), através de regressão linear, chegou-se a uma função linear associando a avaliação da CAPES aos maiores autovalores obtidos.

Outra novidade foi apresentar formas de usar o coeficiente de Gini para análise de grupos de pesquisadores em Redes Sociais Acadêmicas. Especificamente, foram propostas duas diferentes formas de aplicar o coeficiente de Gini, sumarizadas como segue. A primeira proposição foi usar o coeficiente de Gini para avaliar redes ponderadas. As métricas comumente usadas para análise de redes sociais normalmente não consideram os pesos dos relacionamentos, se eles existirem, entre os atores. Foi obtido sucesso em empregar o coeficiente de Gini para medir a homogeneidade do nível de colaboração. A segunda proposição foi usar o coeficiente de Gini combinado com a média de colaboração para ranquear grupos de pesquisa, com base no comportamento colaborativo de seus pesquisadores membros. A hipótese foi de que os melhores grupos de pesquisa são aqueles com a maioria dos pesquisadores contribuindo para a rede do grupo com colaborações, enquanto que os piores grupos são aqueles com somente poucos pesquisadores com algum nível de colaboração. Também foi mostrado que apenas as medidas de média ρ e Gini g_c sozinhas não são suficientes para ranquear grupos de pesquisa. Entretanto, quando combinadas, as duas podem levar a um importante indicador de qualidade para propósitos de ranqueamento (índice β).

Com a análise de diferentes abordagens para funções de recomendação, discutiu-se a falta de abordagens propostas para recomendar colaborações no contexto acadêmico, considerando o perfil do usuário alvo de recomendação (a maioria das abordagens de recomendação de colaboradores focaliza na recomendação de especialistas). Esta tese também apresentou uma abordagem de recomendação de colaborações que trabalhou com a hipótese de que considerar uma similaridade de perfis de pesquisadores (dados por suas áreas de pesquisa) e sua proximidade social pode levar a melhorias significativas na função de recomendação. As novas contribuições podem ser sumarizadas de modo a explorar tal hipótese com base em uma rede social acadêmica. Foi desenvolvido um arcabouço (*framework*) (**CORALS**) para recomendação de colaborações, considerando informações de publicações para construir a rede social. A rede é construída com base em dados extraídos de uma biblioteca digital. Foram analisados diferentes indicadores a serem utilizados no contexto da recomendação de colaboradores (cooperação, correlação e proximidade social) e foram apresentadas métricas para estimar seus valores. Além disso, foi definida uma nova função de recomendação que combina tais indicadores a fim de recomendar colaborações. Essa função também possui um novo tipo de resultado o qual recomenda a intensificação de colaborações já existentes, o que nunca havia sido feito anteriormente.

Foi desenvolvida uma avaliação experimental extensiva utilizando um conjunto de dados real e apresentado um detalhado estudo da influência dos aspectos relacionados a conteúdo (correlação) e questões sociais (cooperação e proximidade social) para gerar recomendações acuradas no contexto acadêmico. A fim de mostrar sua viabilidade, foi construída uma rede atual de pesquisadores (membros de programas de Pós-graduação em Ciência da Computação brasileiros) considerando suas publicações disponibilizadas em uma biblioteca digital conhecida. A qualidade da abordagem de recomendação apresentada nesta tese foi também avaliada contra quatro outras abordagens tradicionais relatadas. Em tal comparação, os valores de revocação obtidos pela abordagem apresentada nesta tese foram maiores do que 9%, uma diferença estatisticamente significativa. Além disso, os resultados mostram que a proximidade social é, na verdade, um indicador muito

relevante para estabelecer novas colaborações. Os resultados também mostraram que a abordagem de recomendação, apresentada nesta tese, melhora os valores de precisão média e de precisão até 10, com resultados estatisticamente significativos, comprovados pelo uso de um teste T. De forma geral, a avaliação experimental também mostrou evidências da importância e utilidade do desenvolvimento de sistemas de recomendação no contexto acadêmico.

Esta tese incluiu propostas para diferentes formas de considerar aspectos temporais na ponderação de colaborações em redes sociais acadêmicas, que são usadas como base para recomendação de conexões. O objetivo final foi analisar a influência desses métodos de ponderação na recomendação de colaboradores no contexto acadêmico. É importante destacar que não foram encontrados na literatura trabalhos correlatos. Também foram apresentados experimentos com conjuntos de dados reais os quais mostraram que considerar aspectos temporais pode conduzir a melhorias na ordenação dos resultados de recomendação. Mais ainda, os resultados mostraram que é possível utilizar aspectos temporais para reduzir o número de relacionamentos considerados para gerar recomendações. O objetivo é que, dessa forma, não sejam considerados relacionamentos que levariam a piorar os resultados de recomendação, por exemplo, relacionamentos pouco frequentes e que acabaram sendo descontinuados.

A seguir, serão apresentadas as produções científicas e trabalhos desenvolvidos relativos à presente tese.

Como resultado da pesquisa bibliográfica inicialmente desenvolvida e relativa aos temas principais deste trabalho, foram publicados dois relatórios técnicos. O primeiro sobre avaliação de qualidade no contexto acadêmico e o segundo relativo a sistemas de recomendação e redes sociais.

- LOPES, Giseli Rabello. **Métricas para avaliação da qualidade de pesquisadores e produções científicas**. Porto Alegre, RS: PPGC da UFRGS, 2008 (Trabalho Individual II (*TI - 1331*)).

- LOPES, Giseli Rabello. **Sistemas de Recomendação e Redes Sociais**. Porto Alegre, RS: PPGC da UFRGS, 2008 (Trabalho Individual III (*TI - 1332*)).

Durante o desenvolvimento desta tese, relativo à proposta para avaliação de qualidade no contexto acadêmico e ao uso do coeficiente de Gini para análise de redes sociais (apresentados no Capítulo 3), foram produzidos os seguintes artigos:

- LOPES, Giseli Rabello; MORO, Mirella M.; SILVA, Roberto da; BARBOSA, Eduardo Martins; OLIVEIRA, José Palazzo Moreira de. **Ranking Strategy for Graduate Programs Evaluation**. In: *The 7th International Conference on Information Technology and Application (ICITA 2011)*, 2011, Sydney, Australia. Proceedings of The 7th International Conference on Information Technology and Application (ICITA 2011), 2011. p. 59-64. (LOPES et al., 2011)

- LOPES, Giseli Rabello; SILVA, Roberto da; OLIVEIRA, José Palazzo Moreira de. **Applying Gini Coefficient to quantify Scientific Collaboration in Researchers Network**. In: *International Workshop on Social Data Mining for Human Behaviour Analysis - SoDaMin*, 2011, Sogndal, Norway. Proceedings of the International Conference on Web Intelligence, Mining and Semantics. New York, NY, USA : ACM, 2011. p. 1-6. (LOPES; SILVA; OLIVEIRA, 2011)

- LOPES, Giseli Rabello; SILVA, Roberto da; MORO, Mirella M.; OLIVEIRA, José Palazzo Moreira de. **Scientific Collaboration in Research Networks: a Quantification Method by using Gini Coefficient**. *IJCSA (International Journal of Computer Science & Applications)*. ISSN 0972-9038. v. 9, n. 2, p. 15-31, 2012. (LOPES et al., 2012)

- LOPES, Giseli Rabello; BRANDÃO, Michele Amaral; BARBOSA, Eduardo M.; MORO, Mirella M.; SILVA, Roberto da; OLIVEIRA, José Palazzo Moreira de. **ReGRaS - Research Group Ranking based on Social Network Analysis**. Periódico: *Informetrics (Journal of Informetrics - Elsevier)*. ISSN: 1751-1577. (A ser submetido)

Adicionalmente, relativo à proposta de recomendação de colaborações em redes sociais acadêmicas e aos refinamentos considerando aspectos temporais (apresentados no Capítulo 4), foram desenvolvidos os seguintes artigos:

- LOPES, Giseli Rabello; MORO, Mirella M.; OLIVEIRA, José Palazzo Moreira de. **Temporal Influence in Collaborators Recommendation on Social Networks**. In: *IADIS International Conference WWW/Internet*, 2011, Rio de Janeiro. Proceedings of the IADIS International Conference WWW/Internet 2011. Lisbon, Portugal : IADIS - International Association for Development of the Information Society, 2011. p. 179-186. (LOPES; MORO; OLIVEIRA, 2011)

- LOPES, Giseli Rabello; MORO, Mirella M.; WIVES, Leandro Krug; OLIVEIRA, José Palazzo Moreira de. **Collaboration Recommendation on Academic Social Networks**. In: *WISM - International Workshop on Web Information Systems Modeling*, 2010, Vancouver. ER 2010 Workshops - LNCS. Berlin, Heidelberg : Springer-Verlag, 2010. v. 6413. p. 190-199. (LOPES et al., 2010a)

- LOPES, Giseli Rabello; MORO, Mirella M.; WIVES, Leandro Krug; OLIVEIRA, José Palazzo Moreira de. **Cooperative Authorship Social Network**. In: *AMW - IV Alberto Mendelzon Workshop on Foundations of Data Management*, 2010, Buenos Aires. Proceedings of the IV Alberto Mendelzon Workshop on Foundations of Data Management, 2010. p. 1.1-1.12. (LOPES et al., 2010b)

- BARBOSA, Eduardo Martins; MORO, Mirella M.; LOPES, Giseli Rabello; OLIVEIRA, José Palazzo Moreira de. **VRRC: Uma Ferramenta Web para Visualização e Recomendação em Redes de Coautoria**. In: *VIII Sessão de Demos, Simpósio Brasileiro de Banco de Dados (SBBD)*, 2011, Florianópolis. Anais do Simpósio Brasileiro de Banco de Dados, 2011. (BARBOSA et al., 2011)

- BARBOSA, Eduardo Martins; MORO, Mirella M.; LOPES, Giseli Rabello; OLIVEIRA, José Palazzo Moreira de. **VRRC: Web Based Tool for Visualization and Recommendation on Co-authorship Network**. In: *ACM SIGMOD International Conference on Management of Data*, 2012, Scottsdale. Proceedings of the International Conference on Management of Data Proceedings, 2012. p. 865-865. (BARBOSA et al., 2012)

- LOPES, Giseli Rabello; MORO, Mirella M.; WIVES, Leandro Krug; OLIVEIRA, José Palazzo Moreira de. **CORALS: a COllaboration Recommender for Academic sociaL networkS**. Periódico: *WWWJ (World Wide Web Journal - Springer)*.

ISSN: 1386-145X (print version) ISSN: 1573-1413 (electronic version). Special Issue - Social Networks and Social Web Mining. (Submetido em jul 2011, Primeira etapa de revisões em nov 2011, Submetida versão revisada em jan 2012)

Além dos citados anteriormente, o artigo a seguir foi publicado, mostrando um apanhado geral dos trabalhos desenvolvidos em ambas as áreas:

- OLIVEIRA, José Palazzo Moreira de; LOPES, Giseli Rabello; MORO, Mirella M. **Academic Social Networks**. In: *8th International Workshop on Web Information Systems Modeling (WISM 2011)*, 2011, Brussels. ER 2011 Workshops - LNCS. Berlin, Heidelberg : Springer-Verlag, 2011. v. 6999. p. 2-3. (OLIVEIRA; LOPES; MORO, 2011)

Também foi desenvolvida a seguinte ferramenta relacionada a esta tese:

- A ferramenta *VRRC* (*V*isualização e *R*ecomendação em *R*edes de *C*oautoria)[1] foi desenvolvida para geração de visualizações em redes sociais acadêmicas, construídas com base em dados sobre publicações. A ferramenta também implementa parte da função de recomendação proposta na presente tese e inclui uma opção de apresentação visual das recomendações geradas. Tal ferramenta recebeu menção honrosa na VIII Sessão de Demos do Simpósio Brasileiro de Banco de Dados (2011).

Essa ferramenta foi desenvolvida como parte do seguinte trabalho de conclusão de curso de graduação:

- Eduardo Martins Barbosa. Visualização de Grafos de Coautoria. 2011. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade Federal de Minas Gerais, Conselho Nacional de Desenvolvimento Científico e Tecnológico. Orientadora: Prof[a]. Dr[a]. Mirella M. Moro.

Além disso, outros trabalhos estão em desenvolvimento explorando assuntos identificados como desdobramentos desta tese, dentre eles:

- Está sendo desenvolvida, junto ao PPGCC da UFMG, sob a orientação da Prof[a]. Dr[a]. Mirella M. Moro, a dissertação de mestrado da aluna Michele Amaral Brandão. Essa dissertação visa a explorar a consideração de aspectos de localização geográfica na recomendação de colaborações no contexto acadêmico.

5.2 Trabalhos Futuros

A seguir, são mostrados alguns resultados que foram obtidos para: enfatizar a validade de considerar as colaborações, em redes acadêmicas, para inferir aspectos de qualidade de grupos e recomendar colaborações; e orientar a discussão vislumbrando trabalhos futuros. O conjunto de dados usado inclui os 732 pesquisadores de 27 programas de Pós-graduação em Computação brasileiros (mesmos da Tabela 3.3) e suas publicações (que foram extraídas da DBLP em Julho de 2011). O objetivo é analisar o ano limite da última classificação da CAPES (2009) e o período posterior a essa avaliação (2010-2011).

[1] VRRC: http://www.lbd.dcc.ufmg.br/vrrc/

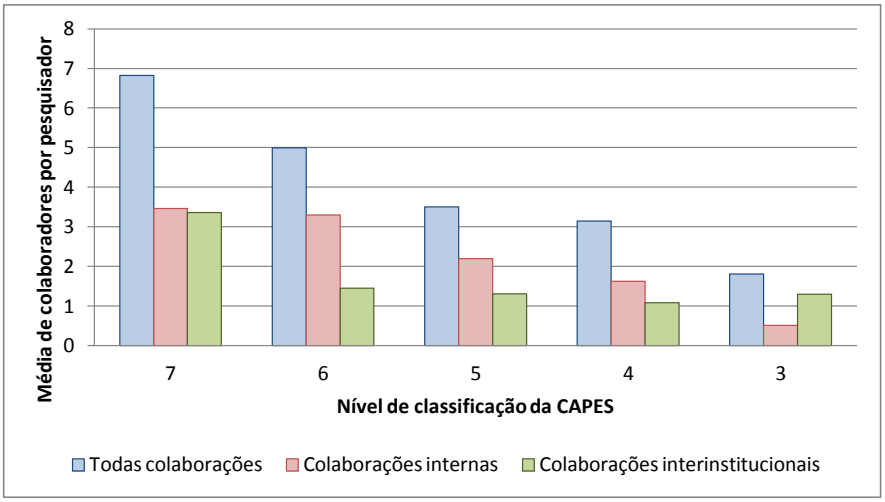


Figura 5.1: Média de colaboradores por pesquisador agrupada por nível de classificação da CAPES, considerando publicações até o ano de 2009.

A Figura 5.1 apresenta os resultados para a média de colaboradores por pesquisador pelo nível CAPES, considerando as publicações até o ano de 2009. A Figura 5.2 apresenta os resultados, considerando somente os novos colaboradores, que nunca colaboraram juntos antes, encontrados em artigos publicados depois do ano de 2009. Os dois gráficos são construídos, considerando somente os colaboradores que são membros dos 27 programas de Pós-graduação escolhidos e eles apresentam os resultados para: (i) todas as colaborações representadas na SN, (ii) colaborações internas e (iii) colaborações interinstitucionais. Os resultados em ambos os gráficos mostraram que em geral a média do número de colaboradores por pesquisador para grupos de níveis altos é maior do que para grupos de níveis mais baixos. Isso mostra evidências de que quanto maior o nível (qualidade do programa), maior a tendência de comportamento colaborativo. A única exceção, em alguns casos, ocorre no nível 3 (para o qual apenas uma pequena porção de todos os programas foi considerada; escolheu-se aqueles que possuíam alguns de seus pesquisadores com publicações indexadas na DBLP). Isso provavelmente ocorreu porque alguns membros de diversos programas de Pós-graduação de nível 3 não tinham publicações indexadas na DBLP e a média obtida, utilizando os programas escolhidos, pode não representar adequadamente o comportamento colaborativo médio de todos os programas desse nível.

Os resultados apresentados na Figura 5.1 mostram evidências de que analisar o comportamento colaborativo pode ser importante para indicar a qualidade do grupo. Mais ainda, os resultados apresentados na Figura 5.2 indicam que o comportamento colaborativo continua sendo verificado nos programas de nível alto também após 2009, com a busca de colaborações com novos pesquisadores.

Assim, uma abordagem, considerando o aspecto de qualidade inferida com base em colaborações prévias, pode ser usada para melhorar os resultados de recomendação. Na Figura 5.1 pode-se observar que pesquisadores membros de programas de alta qualidade estão mais "abertos" para iniciar novas colaborações com outros pesquisadores (incluindo colaborações internas e interinstitucionais). Dessa forma, a recomendação de pesquisadores membros de programas de nível mais elevado, para pesquisadores membros de programas de níveis mais baixos, pode ser muito valorosa para melhorar a inserção de programas de Pós-graduação emergentes (com comportamento colaborativo baixo quando

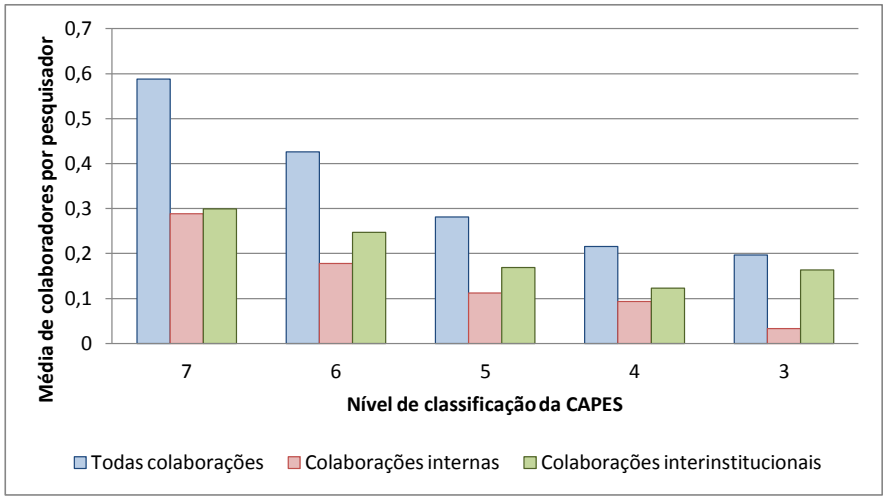


Figura 5.2: Média de novos colaboradores por pesquisador agrupada por nível de classificação da CAPES considerando coautorias iniciadas após o ano de 2009.

comparado com outros programas de nível mais elevado) no cenário de pesquisas nacionais e internacionais. Além disso, um ponto que começou a ser pesquisado e que a visa ser aperfeiçoado em trabalhos futuros (até pela exigência de avaliações longas com diversos usuários e grupos de pesquisa distintos - com e sem intervenção) é utilizar a avaliação de grupos de pesquisadores para ajudar a refinar ainda mais a abordagem de recomendação. Uma alternativa seria determinar os tipos de recomendação que devem ser priorizados. Por exemplo, em grupos fracamente conectados, a opção de priorizar a recomendação de maior intensificação entre seus membros com fracas conexões e também a recomendação de novas colaborações dentro do próprio grupo, visando uma melhora interna antes de se passar para um passo maior de interação com pesquisadores de outras instituições. Por outro lado, para grupos altamente conectados internamente, o ideal é recomendar principalmente intensificações de colaborações com membros externos e a definição de novas colaborações também com membros externos. Isso pensado em um nível específico, dentro de uma instituição, mas pode ser expandido para um nível mais global, em relação a cooperações interinstitucionais e até internacionais. A ideia a ser explorada é relacionada aos benefícios da recomendação tanto para os indivíduos como para os grupos de pesquisa interna e externamente.

Alguns trabalhos futuros identificados, especificamente em relação a extensões na abordagem de avaliação de qualidade, são complementar o estudo das colaborações e considerar também os indivíduos externos ao programa de Pós-graduação sendo avaliado. Por exemplo, alguns indivíduos identificados como “ineficientes sociais” podem ter comportamento colaborativo com pesquisadores externos.

Como trabalhos futuros em relação à abordagem de recomendação, pretende-se trabalhar na hipótese de considerar-se outros indicadores além da cooperação, correlação e proximidade social. Além disso, pretende-se ampliar ainda mais o estudo considerando a influência de aspectos temporais. Um dos próximos passos é analisar essa faceta, incluindo mais dados disponibilizados em outras bibliotecas digitais (além da DBLP) e mídias sociais. Como já mencionado nas Seções 4.2.4 e 4.5.2.2, também se planeja considerar o *feedback* dos usuários, a fim de, futuramente, otimizar ainda mais a função de recomendação. Nos experimentos apresentados, para compor a rede social de possíveis recomendáveis, foram selecionados os pesquisadores que eram membros de certos pro-

gramas de Pós-graduação. Nesse sentido, futuramente, visa-se definir critérios de seleção do subgrupo a ser recomendado, com o uso das métricas associadas aos indicadores de qualidade de grupos, por exemplo, utilizar um ponto de corte para recomendar somente pesquisadores associados a instituições classificadas acima de determinado "nível" de qualidade.

Especificamente sobre os refinamentos temporais, também pode-se estudar a influência dos aspectos temporais quando são utilizadas outras funções de escore (outros métodos de predição de ligações além do menor caminho). Além disso, as melhorias, obtidas pelo uso de aspectos temporais e das métricas definidas, planejam ser testadas em outras redes sociais, por exemplo, em redes de relacionamento, para conexão de amigos. Também relativo aos aspectos temporais, foi inicialmente explorado seu uso para redução dos relacionamentos utilizados na rede social analisada para recomendação. Futuramente, pretende-se explorar essa faceta na avaliação de qualidade. Outro exemplo de aplicação seria para determinação de uma componente principal, ou mais importante, a ser analisada em uma rede global para gerar recomendações a um determinado usuário (até por questões de escalabilidade).

Além disso, está em desenvolvimento um novo projeto de cooperação internacional. Nesse projeto, serão aplicados os mecanismos propostos nesta tese de avaliação e recomendação para analisar os grupos de pesquisa dos diferentes países membros.

Os resultados obtidos pela presente tese apontaram indícios da validade e aplicabilidade de suas proposições e apontam possibilidades de novas pesquisas para sua continuidade e extensão.

REFERÊNCIAS

ADLER, R.; EWING, J.; TAYLOR, P. **Citation Statistics**. [S.l.: s.n.], 2008. Technical Report, Disponível em: <http://www.mathunion.org/Publications/Report/CitationStatistics>. Acesso em: dez. 2011.

ADOMAVICIUS, G.; TUZHILIN, A. Toward the Next Generation of Recommender Systems: a survey of the state-of-the-art and possible extensions. **IEEE Trans. on Knowl. and Data Eng.**, Piscataway, NJ, USA, v.17, n.6, p.734–749, 2005.

ALEMAN-MEZA, B.; NAGARAJAN, M.; RAMAKRISHNAN, C.; DING, L.; KOLARI, P.; SHETH, A. P.; ARPINAR, I. B.; JOSHI, A.; FININ, T. Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection. In: WORLD WIDE WEB, WWW '06, 15., 2006, New York, NY, USA. **Proceedings...** ACM, 2006. p.407–416.

ANGELOVA, R.; KASNECI, G.; WEIKUM, G. Graffiti: graph-based classification in heterogeneous networks. **World Wide Web**, [S.l.], 2011. Published online first.

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. A. **Modern Information Retrieval - the concepts and technology behind search, Second edition**. [S.l.]: Pearson Education Ltd., Harlow, England, 2011.

BALABANOVIC, M.; SHOHAM, Y. Content-Based, Collaborative Recommendation. **Commun. ACM**, [S.l.], v.40, n.3, p.66–72, 1997.

BARABÁSI, A.-L. **Linked**: the new science of networks. [S.l.]: Basic Books, 2002.

BARBOSA, E. M.; MORO, M. M.; LOPES, G. R.; OLIVEIRA, J. P. M. de. VRRC: uma ferramenta web para visualização e recomendação em redes de coautoria. In: VIII SESSÃO DE DEMOS, SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBD), 2011. **Proceedings...** [S.l.: s.n.], 2011.

BARBOSA, E. M.; MORO, M. M.; LOPES, G. R.; OLIVEIRA, J. P. M. de. VRRC: web based tool for visualization and recommendation on co-authorship network. In: INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2012, New York, NY, USA. **Proceedings...** ACM, 2012. p.865–865. (SIGMOD '12).

BARNES, J. A. Class and Committees in a Norwegian Island Parish. **Human Relations**, [S.l.], v.7, n.1, p.39–58, Feb. 1954.

BERGAMASCHI, S.; GUERRA, F.; LEIBA, B. Guest Editors' Introduction: information overload. **Internet Computing, IEEE**, [S.l.], v.14, n.6, p.10 –13, 2010.

BERNERS-LEE, T.; FISCHETTI, M. **Weaving the Web** : the original design and ultimate destiny of the world wide web by its inventor. [S.l.]: Harper San Francisco, 1999.

BILENKO, M.; MOONEY, R. J. Adaptive duplicate detection using learnable string similarity measures. In: ACM SIGKDD INTL. CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING, 2003, Washington, DC. **Proceedings...** [S.l.: s.n.], 2003. p.39–48.

BORGES, E. N.; CARVALHO, M. G. de; GALANTE, R.; GONÇALVES, M. A.; LAENDER, A. H. F. An unsupervised heuristic-based approach for bibliographic metadata deduplication. **Inf. Process. Manage.**, Tarrytown, NY, USA, v.47, p.706–718, September 2011.

BUCKLEY, C.; VOORHEES, E. M. Evaluating evaluation measure stability. In: ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 23., 2000, New York, NY, USA. **Proceedings...** ACM, 2000. p.33–40. (SIGIR '00).

BURKE, R. D. Hybrid Recommender Systems: survey and experiments. **User Model. User-Adapt. Interact.**, [S.l.], v.12, n.4, p.331–370, 2002.

CARVALHO, M. G.; LAENDER, A. H. F.; GONÇALVES, M. A.; SILVA, A. S. da. Replica identification using genetic programming. In: ACM SYMPOSIUM ON APPLIED COMPUTING, 2008, Fortaleza, Brazil. **Proceedings...** [S.l.: s.n.], 2008. p.1801–1806.

CHEN, H.-H.; GOU, L.; ZHANG, X.; GILES, C. L. CollabSeer: a search engine for collaboration discovery. In: ACM/IEEE JOINT CONFERENCE ON DIGITAL LIBRARIES, 11., 2011, New York, NY, USA. **Proceedings...** ACM, 2011. p.231–240. (JCDL '11).

CHEN, J.; GEYER, W.; DUGAN, C.; MULLER, M.; GUY, I. Make new friends, but keep the old: recommending people on social networking sites. In: HUMAN FACTORS IN COMPUTING SYSTEMS, CHI '09, 27., 2009, New York, NY, USA. **Proceedings...** ACM, 2009. p.201–210.

CHERKASSKY, B. V.; GOLDBERG, A. V.; RADZIK, T. Shortest paths algorithms: theory and experimental evaluation. **Math. Program.**, Secaucus, NJ, USA, v.73, n.2, p.129–174, 1996.

CLAYPOOL, M.; GOKHALE, A.; MIRANDA, T.; MURNIKOV, P.; NETES, D.; SARTIN, M. Combining Content-Based and Collaborative Filters in an Online Newspaper. In: ACM SIGIR WORKSHOP ON RECOMMENDER SYSTEMS, 1999, Berkley, California. **Proceedings...** New York: ACM Press, 1999.

COHEN, W. W.; RICHMAN, J. Learning to match and cluster large high-dimensional data sets for data integration. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2002, New York, NY, USA. **Proceedings...** ACM, 2002. p.475–480. (KDD '02).

COTA, R. G.; FERREIRA, A. A.; NASCIMENTO, C.; GONÇALVES, M. A.; LAENDER, A. H. F. An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. **JASIST**, [S.l.], v.61, n.9, p.1853–1870, 2010.

DA LUZ, M.; MARQUES-PORTELLA, C.; MENDLOWICZ, M.; GLEISER, S.; SILVA FREIRE COUTINHO, E.; FIGUEIRA, I. Institutional h-index: the performance of a new metric in the evaluation of brazilian psychiatric post-graduation programs. **Scientometrics**, [S.l.], v.77, p.361–368, 2008.

DING, Y. Scientific collaboration and endorsement: network analysis of coauthorship and citation networks. **Journal of Informetrics**, [S.l.], v.5, n.1, p.187 – 203, 2011.

EASLEY, D.; KLEINBERG, J. **Networks, Crowds, and Markets**: reasoning about a highly connected world. New York, NY, USA: Cambridge University Press, 2010.

EGGHE, L. An improvement of the h-index: the g-index. **ISSI Newsletter**, [S.l.], v.2, n.1, p.8–9, 2006.

EGGHE, L. The Hirsch-index and related impact measures. **Annual Review of Information Science and Technology**, [S.l.], v.44, p.65–114, 2010.

FIGUEIRA FILHO, F. M.; ALBUQUERQUE, J. Porto de; GEUS, P. de. Sistemas de recomendação e interação na web social. In: WORKSHOP ON HUMAN-COMPUTER INTERACTION ASPECTS IN THE SOCIAL WEB, IN CONJUNCTION WITH THE VIII BRAZILIAN SYMPOSIUM OF HUMAN FACTORS ON COMPUTER SYSTEMS (IHC'08), 1., 2008. **Proceedings...** [S.l.: s.n.], 2008. p.24–27.

FREEMAN, L. The impact of computer based communication on the social structure of an emerging scientific specialty. **Social Networks 6**, [S.l.], p.201–221, 1984.

FREEMAN, L. C. Centrality in social networks: conceptual clarification. **Social Networks**, [S.l.], v.1, p.215–239, 1979.

FREIRE, V.; FIGUEIREDO, D. R. Ranqueamento em Redes de Colaboração Utilizando uma Métrica Baseada em Intensidade do Relacionamento. In: SIMPÓSIO BRASILEIRO DE SISTEMAS COLABORATIVOS (SBSC), 2010. **Anais...** [S.l.: s.n.], 2010.

GARFIELD, E. Citation indexes for science: a new dimension in documentation through association of ideas. **Science**, [S.l.], v.122, p.108–111, 1955.

GARFIELD, E. The history and meaning of the journal impact factor. **JAMA**, Thomson Scientific, Philadelphia, USA, v.295, n.1, p.90–93, Jan. 2006.

GEYER, W.; DUGAN, C.; MILLEN, D. R.; MULLER, M.; FREYNE, J. Recommending topics for self-descriptions in online user profiles. In: ACM CONFERENCE ON RECOMMENDER SYSTEMS, RECSYS '08, 2008., 2008, New York, NY, USA. **Proceedings...** ACM, 2008. p.59–66.

GINI, C. W. Variability and Mutability, contribution to the study of statistical distributions and relations. **Studi Economico-Giuridici della R. Universita de Cagliari**, [S.l.], 1912. Reviewed in: Light, R.J., Margolin, B.H.: An Analysis of Variance for Categorical Data. *J. American Statistical Association*, Vol. 66 pp. 534-544 (1971).

GIULIANI, F.; DE PETRIS, M.; NICO, G. Assessing scientific collaboration through coauthorship and content sharing. **Scientometrics**, [S.l.], v.85, n.1, p.13–28, Oct. 2010.

GOLBECK, J.; HENDLER, J. FilmTrust: movie recommendations using trust in web-based social networks. In: CONSUMER COMMUNICATIONS AND NETWORKING CONFERENCE, 2006. CCNC 2006. 3RD IEEE, 2006. **Proceedings...** [S.l.: s.n.], 2006. v.1, p.282–286.

GOLDBERG, D.; NICHOLS, D.; OKI, B. M.; TERRY, D. Using collaborative filtering to weave an information tapestry. **Commun. ACM**, New York, NY, USA, v.35, n.12, p.61–70, 1992.

GOWRISHANKAR, J.; DIVAKAR, P.; BAYLIS, M.; GRAVENOR, M.; KAO, R. Sprucing up one's impact factor. **Nature**, [S.l.], v.401, p.321–322, Sept. 1999.

GUY, I.; ZWERDLING, N.; RONEN, I.; CARMEL, D.; UZIEL, E. Social media recommendation based on people and tags. In: ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 33., 2010, New York, NY, USA. **Proceedings...** ACM, 2010. p.194–201. (SIGIR '10).

HABIBZADEH, F.; YADOLLAHIE, M. Journal weighted impact factor: a proposal. **Journal of Informetrics**, [S.l.], v.2, n.2, p.164–172, Apr. 2008.

HAN, H.; GILES, L.; ZHA, H.; LI, C.; TSIOUTSIOULIKLIS, K. Two supervised learning approaches for name disambiguation in author citations. In: ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES, 4., 2004, New York, NY, USA. **Proceedings...** ACM, 2004. p.296–305. (JCDL '04).

HECK, T.; HANRATHS, O.; STOCK, W. G. Expert Recommendation for Knowledge Management in Academia. In: ASIST, 2011, New Orleans, LA, USA. **Proceedings...** [S.l.: s.n.], 2011. (ASIST '11).

HERLOCKER, J. L. **Understanding and improving automated collaborative filtering systems**. 2000. Doctoral Dissertation — University of Minnesota, Minnesota. Adviser-Joseph A. Konstan.

HIRSCH, J. E. An index to quantify an individual's scientific research output. **Proceedings of the National Academy of Sciences**, [S.l.], v.102, n.46, p.16569–16572, Nov. 2005.

HOSER, B.; HOTHO, A.; JÄSCHKE, R.; SCHMITZ, C.; STUMME, G. Semantic Network Analysis of Ontologies. In: EUROPEAN SEMANTIC WEB CONFERENCE, 3., 2006. **Proceedings...** Springer, 2006. v.4011, p.514–529.

HUANG, Z.; CHUNG, W.; ONG, T.-H.; CHEN, H. A graph-based recommender system for digital library. In: JOINT CONFERENCE ON DIGITAL LIBRARIES, JCDL, 2, 2002, New York, NY, USA. **Proceedings...** ACM Press, 2002. p.65–73.

HULL, D. Using statistical testing in the evaluation of retrieval experiments. In: ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 16., 1993, New York, NY, USA. **Proceedings...** ACM, 1993. p.329–338. (SIGIR '93).

HWANG, S. Y.; HSIUNG, W. C.; YANG, W. S. A prototype WWW literature recommendation system for digital libraries. **Online Information Review**, [S.l.], v.27, n.3, p.169–182, 2003.

HWANG, S.-Y.; WEI, C.-P.; LIAO, Y.-F. Coauthorship networks and academic literature recommendation. **Electron. Commer. Rec. Appl.**, Amsterdam, The Netherlands, The Netherlands, v.9, n.4, p.323–334, 2010.

KARAGIANNIS, T.; VOJNOVIC, M. Behavioral profiles for advanced email features. In: WORLD WIDE WEB, WWW '09, 18., 2009, New York, NY, USA. **Proceedings...** ACM, 2009. p.711–720.

KATZ, L. A new status index derived from sociometric analysis. **Psychometrika**, [S.l.], v.18, n.1, p.39–43, March 1953.

KAUTZ, H.; SELMAN, B.; SHAH, M. Referral Web: combining social networks and collaborative filtering. **Commun. ACM**, New York, NY, USA, v.40, n.3, p.63–65, 1997.

KNOKE, D.; YANG, S. **Social Network Analysis**. 2.ed. [S.l.]: Sage Publications, Inc, 2007. 144p. (Quantitative Applications in the Social Sciences).

KONSTAN, J. A.; MILLER, B. N.; MALTZ, D.; HERLOCKER, J. L.; GORDON, L. R.; RIEDL, J. GroupLens: applying collaborative filtering to usenet news. **Commun. ACM**, New York, NY, USA, v.40, n.3, p.77–87, 1997.

LAENDER, A. H. F.; LUCENA, C. J. P. de; MALDONADO, J. C.; SILVA, E. de Souza e; ZIVIANI, N. Assessing the research and education quality of the top Brazilian Computer Science graduate programs. **SIGCSE Bull.**, New York, NY, USA, v.40, p.135–145, June 2008.

LAENDER, A. H. F.; MORO, M. M.; SILVA, A. S. da; JR., C. A. D.; GONÇALVES, M. A.; GALANTE, R.; SILVA, A. J. C.; BIGONHA, C. A. S.; DALIP, D. H.; BARBOSA, E. M.; BORGES, E. N.; CORTEZ, E.; JR., P. S. P.; ALENCAR, R. O. de; CARDOSO, T. N. C.; SALLES, T. CiênciaBrasil - The Brazilian Portal of Science and Technology. In: SEMISH - SEMINÁRIO INTEGRADO DE SOFTWARE E HARDWARE, 2011. **Proceedings...** [S.l.: s.n.], 2011.

LANG, K. NewsWeeder: learning to filter netnews. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 12., 1995. **Proceedings...** Morgan Kaufmann publishers Inc.: San Mateo: CA: USA, 1995. p.331–339.

LI, L.; OTSUKA, S.; KITSUREGAWA, M. Finding Related Search Engine Queries by Web Community Based Query Enrichment. **World Wide Web**, [S.l.], v.13, n.1, p.121–142, 2010.

LIBEN-NOWELL, D.; KLEINBERG, J. The link-prediction problem for social networks. **J. Am. Soc. Inf. Sci. Technol.**, New York, NY, USA, v.58, n.7, p.1019–1031, 2007.

LIEBERMAN, H. Autonomous interface agents. In: SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, CHI '97, 1997, New York, NY, USA. **Proceedings...** ACM, 1997. p.67–74.

LINDEN, G.; SMITH, B.; YORK, J. Amazon.com recommendations: item-to-item collaborative filtering. **Internet Computing, IEEE**, [S.l.], v.7, n.1, p.76 – 80, 2003.

LIU, X.; BOLLEN, J.; NELSON, M. L.; SOMPEL, H. Van de. Co-authorship networks in the digital library research community. **Inf. Process. Manage.**, [S.l.], v.41, n.6, p.1462–1480, 2005.

LOH, S.; LICHTNOW, D.; BORGES, T.; PILTCHER, G.; NUNES, M. F. Constructing domain ontologies for indexing texts and creating users'profiles. In: WORKSHOP ON ONTOLOGIES AND METAMODELING IN SOFTWARE AND DATA ENGINEERING - WOMSDE, SBBD'2006, 1., 2006, Florianópolis, SC, Brazil. **Proceedings...** [S.l.: s.n.], 2006. p.72–82.

LOH, S.; LICHTNOW, D.; KAMPFF, A. J.; OLIVEIRA, J. P. M. d. Recommendation of Complementary Material during Chat Discussions. **Knowledge Management**, [S.l.], v.2, n.4, p.385–399, 2010.

LOH, S.; WIVES, L. K.; OLIVEIRA, J. P. M. d. Concept-based knowledge discovery in texts extracted from the Web. **SIGKDD Explor. Newsl.**, New York, NY, USA, v.2, n.1, p.29–39, June 2000.

LOPES, G. R.; MORO, M. M.; OLIVEIRA, J. P. M. de. Temporal Influence in Collaborators Recommendation in Social Networks. In: IADIS INTERNATIONAL CONFERENCE WWW/INTERNET, 2011, Lisbon, Portugal. **Proceedings...** IADIS - International Association for Development of the Information Society, 2011. p.179–186.

LOPES, G. R.; MORO, M. M.; SILVA, R. da; BARBOSA, E. M.; OLIVEIRA, J. P. M. de. Ranking Strategy for Graduate Programs Evaluation. In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY AND APPLICATION (ICITA 2011), 7., 2011. **Proceedings...** [S.l.: s.n.], 2011. p.69–74.

LOPES, G. R.; MORO, M. M.; WIVES, L. K.; OLIVEIRA, J. P. M. d. Collaboration Recommendation on Academic Social Networks. In: TRUJILLO, J.; DOBBIE, G.; KANGASSALO, H.; HARTMANN, S.; KIRCHBERG, M.; ROSSI, M.; REINHARTZBERGER, I.; ZIMáNYI, E.; FRASINCAR, F. (Ed.). **Advances in Conceptual Modeling - Applications and Challenges**. [S.l.]: Springer Berlin / Heidelberg, 2010. p.190–199. (Lecture Notes in Computer Science, v.6413). 10.1007/978-3-642-16385-2_24.

LOPES, G. R.; MORO, M. M.; WIVES, L. K.; OLIVEIRA, J. P. M. de. Cooperative Authorship Social Network. In: IV ALBERTO MENDELZON WORKSHOP ON FOUNDATIONS OF DATA MANAGEMENT (AMW 2010), 2010, Buenos Aires, Argentina. **Proceedings...** CEUR-WS.org, 2010. p.1:1–1:12. (CEUR Workshop Proceedings, v.619).

LOPES, G. R.; SILVA, R. da; MORO, M. M.; OLIVEIRA, J. P. M. de. Scientific Collaboration in Research Networks: a quantification method by using gini coefficient. **IJCSA**, [S.l.], v.9, n.2, p.15–31, 2012.

LOPES, G. R.; SILVA, R. da; OLIVEIRA, J. P. M. de. Applying Gini coefficient to quantify scientific collaboration in researchers network. In: INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, MINING AND SEMANTICS, INTERNATIONAL WORKSHOP ON SOCIAL DATA MINING FOR HUMAN BEHAVIOUR ANALYSIS

- SODAMIN, 2011, New York, NY, USA. **Proceedings...** ACM, 2011. p.68:1–68:6. (WIMS '11).

LOPES, G. R.; SOUTO, M. A. M.; WIVES, L. K.; OLIVEIRA, J. P. M. d. A personalized recommender system for digital libraries. In: BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, 14., 2008, New York, NY, USA. **Proceedings...** ACM, 2008. p.59–66. (WebMedia '08).

MAEDCHE, A.; STAAB, S. Ontology Learning for the Semantic Web. **IEEE Intelligent Systems**, Piscataway, NJ, USA, v.16, n.2, p.72–79, 2001.

MAES, P. Agents that reduce work and information overload. **Commun. ACM**, New York, NY, USA, v.37, n.7, p.30–40, 1994.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to Information Retrieval**. [S.l.]: Cambridge University Press, 2008.

MARSDEN, P. V. Egocentric and sociocentric measures of network centrality. **Social Networks**, [S.l.], v.24, n.4, p.407–422, 2002.

MCDONALD, D. W. Recommending collaboration with social networks: a comparative evaluation. In: SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, CHI '03, 2003, New York, NY, USA. **Proceedings...** ACM, 2003. p.593–600.

MCSHERRY, F.; NAJORK, M. Computing Information Retrieval Performance Measures Efficiently in the Presence of Tied Scores. In: MACDONALD, C.; OUNIS, I.; PLACHOURAS, V.; RUTHVEN, I.; WHITE, R. (Ed.). **Advances in Information Retrieval**. [S.l.]: Springer Berlin / Heidelberg, 2008. p.414–421. (Lecture Notes in Computer Science, v.4956). 10.1007/978-3-540-78646-7_38.

MENEZES, G. V.; ZIVIANI, N.; LAENDER, A. H.; ALMEIDA, V. A geographical analysis of knowledge production in computer science. In: WORLD WIDE WEB, WWW '09, 18., 2009, New York, NY, USA. **Proceedings...** ACM, 2009. p.1041–1050.

MEO, P. D.; NOCERA, A.; ROSACI, D.; URSINO, D. Recommendation of reliable users, social networks and high-quality resources in a Social Internetworking System. **AI Commun.**, [S.l.], v.24, n.1, p.31–50, 2011.

MIKA, P. Social Networks and the Semantic Web. In: IEEE/WIC/ACM INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, 2004., 2004, Washington, DC, USA. **Proceedings...** IEEE Computer Society, 2004. p.285–291. (WI '04).

MILGRAM, S. The Small World Problem. **Psychology Today**, [S.l.], v.2, p.60–67, 1967.

MOLINARI, J.-F.; MOLINARI, A. A new methodology for ranking scientific institutions. **Scientometrics**, [S.l.], v.75, p.163–174, 2008.

MONTANER, M.; LÓPEZ, B.; ROSA, J. L. de la. A Taxonomy of Recommender Agents on the Internet. **Artif. Intell. Rev.**, [S.l.], v.19, n.4, p.285–330, 2003.

MOONEY, R. J.; ROY, L. Content-based book recommending using learning for text categorization. In: ACM CONFERENCE ON DIGITAL LIBRARIES, DL '00, 2000, New York, NY, USA. **Proceedings...** ACM, 2000. p.195–204.

MORO, M. M.; LIM, L.; CHANG, Y.-C. Schema advisor for hybrid relational-XML DBMS. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2007., 2007, New York, NY, USA. **Proceedings...** ACM, 2007. p.959–970. (SIGMOD '07).

NAJIMINAINI, M.; SUBEDI, L.; TRAJKOVIC, L. Analysis of Internet topologies: a historical view. In: INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS, ISCAS 2009, 2009, Taipei, Taiwan. **Proceedings...** IEEE, 2009. p.1697–1700.

NASCIMENTO, M. A.; SANDER, J.; POUND, J. Analysis of SIGMOD's co-authorship graph. **SIGMOD Rec.**, New York, NY, USA, v.32, n.3, p.8–10, 2003.

NEWMAN, M. E. J. Scientific collaboration networks. I. network construction and fundamental results. **Physical Review E**, [S.l.], v.64, n.1, p.016131, 2001.

NEWMAN, M. E. J. The Structure and Function of Complex Networks. **SIAM Review**, [S.l.], v.45, n.2, p.167–256, 2003.

NICOLAISEN, J.; FRANDSEN, T. F. The Reference Return Ratio. **Journal of Informetrics**, [S.l.], v.2, n.2, p.128–135, Apr. 2008.

OGATA, H.; YANO, Y.; FURUGORI, N.; JIN, Q. Computer Supported Social Networking For Augmenting Cooperation. **Comput. Supported Coop. Work**, Norwell, MA, USA, v.10, n.2, p.189–209, 2001.

OLIVEIRA, J. P. M. de; LOPES, G. R.; MORO, M. M. Academic Social Networks. In: DE TROYER, O.; BAUZER MEDEIROS, C.; BILLEN, R.; HALLOT, P.; SIMITSIS, A.; VAN MINGROOT, H. (Ed.). **Advances in Conceptual Modeling. Recent Developments and New Directions**. [S.l.]: Springer Berlin / Heidelberg, 2011. p.2–3. (Lecture Notes in Computer Science, v.6999).

OLSINA, L.; ANGELES MARTÍN, M. de los. Ontology for Software Metrics and Indicators: building process and decisions taken. In: ICWE - WEB ENGINEERING - 4TH INTERNATIONAL CONFERENCE, 2004, Munich, Germany. **Proceedings...** Springer, 2004. p.176–181. (Lecture Notes in Computer Science, v.3140).

PAZZANI, M. J. A Framework for Collaborative, Content-Based and Demographic Filtering. **Artif. Intell. Rev.**, Norwell, MA, USA, v.13, n.5-6, p.393–408, 1999.

PERUGINI, S.; GONÇALVES, M. A.; FOX, E. A. Recommender Systems Research: a connection-centric survey. **J. Intell. Inf. Syst.**, Hingham, MA, USA, v.23, n.2, p.107–143, 2004.

PISSARD, N.; PRIEUR, C. Thematic vs. social networks in web 2.0 communities: a case study on flickr groups. In: ALGOTEL CONFERENCE, 2007. **Proceedings...** [S.l.: s.n.], 2007.

QUERCIA, D.; CAPRA, L. FriendSensing: recommending friends using mobile phones. In: ACM CONFERENCE ON RECOMMENDER SYSTEMS, RECSYS '09, 2009, New York, NY, USA. **Proceedings...** ACM, 2009. p.273–276.

RAAN, A. F. J. van. Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. **Scientometrics**, [S.l.], v.67, p.491–502, 2006.

RAHM, E.; DO, H. H. Data Cleaning: problems and current approaches. **IEEE Data Eng. Bull.**, [S.l.], v.23, n.4, p.3–13, 2000.

REATEGUI, E. B.; CAZELLA, S. C. Sistemas de Recomendação. **ENIA**, [S.l.], v.V, p.306–348, 2005.

RECUERO, R. Redes Sociais na Internet: considerações iniciais. **E-Compós**, Brasília, v.2, 2005.

REITZ, F.; HOFFMANN, O. An analysis of the evolving coverage of computer science sub-fields in the DBLP digital library. In: EUROPEAN CONFERENCE ON RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES, 14., 2010, Berlin, Heidelberg. **Proceedings...** Springer-Verlag, 2010. p.216–227. (ECDL'10).

REN, J.; TAYLOR, R. N. Automatic and versatile publications ranking for research institutions and scholars. **Commun. ACM**, New York, NY, USA, v.50, n.6, p.81–85, 2007.

RESNICK, P.; IACOVOU, N.; SUCHAK, M.; BERGSTROM, P.; RIEDL, J. Group Lens: an open architecture for collaborative filtering of netnews. In: ACM 1994 CONFERENCE ON COMPUTER SUPPORTED COOPERATIVE WORK, 1994. **Proceedings...** [S.l.: s.n.], 1994. p.175–186.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Inf. Process. Manage.**, Tarrytown, NY, USA, v.24, n.5, p.513–523, 1988.

SCHAFER, J. B.; KONSTAN, J. A.; RIEDL, J. E-Commerce Recommendation Applications. **Data Mining and Knowledge Discovery**, Hingham, MA, USA, v.5, n.1-2, p.115–153, 2001.

SHAHABI, C.; CHEN, Y.-S. An Adaptive Recommendation System without Explicit Acquisition of User Relevance Feedback. **Distrib. Parallel Databases**, Hingham, MA, USA, v.14, n.2, p.173–192, 2003.

SHARDANAND, U.; MAES, P. Social information filtering: algorithms for automating "word of mouth". In: SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, CHI '95, 1995, New York, NY, USA. **Proceedings...** ACM Press/Addison-Wesley Publishing Co., 1995. p.210–217.

SIEGEL, S.; CASTELLAN, N. J. **Nonparametric Statistics for the Behavioral Sciences**. 2.ed. New York: McGraw-Hill, 1988.

SILVA, R. da; KALIL, F.; OLIVEIRA, J. P. M. de; MARTINEZ, A. S. Universality in Bibliometrics. **Physica A**, [S.l.], v.391, n.5, p.2119–2128, 2012.

SILVA, R. da; OLIVEIRA, J. P. M. de; LIMA, J. V. de; MOREIRA, V. Statistics for Ranking Program Committees and Editorial Boards. **CoRR**, [S.l.], v.abs/1002.1060, 2010.

SMEATON, A.; CALLAN, J. Joint DELOS-NSF workshop on personalisation and recommender systems in digital libraries. **SIGIR Forum**, New York, NY, USA, v.35, n.1, p.7–11, 2001.

SMEATON, A. F.; CALLAN, J. Personalisation and recommender systems in digital libraries. **International Journal on Digital Libraries**, [S.l.], v.V5, n.4, p.299–308, 2005.

SMEATON, A. F.; KEOGH, G.; GURRIN, C.; MCDONALD, K.; S/ODRING, T. Analysis of papers from twenty-five years of SIGIR conferences: what have we been doing for the last quarter of a century. **ACM SIGIR Forum**, [S.l.], v.36, n.2, p.49–53, 2003.

SONG, Y.; HUANG, J.; COUNCILL, I. G.; LI, J.; GILES, C. L. Efficient topic-based unsupervised name disambiguation. In: ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES, 7., 2007, New York, NY, USA. **Proceedings...** ACM, 2007. p.342–351. (JCDL '07).

SOUTO, M. A. M.; WARPECHOWSKI, M.; OLIVEIRA, J. P. M. d. An ontological approach for the quality assessment of computer science conferences. In: ADVANCES IN CONCEPTUAL MODELING: FOUNDATIONS AND APPLICATIONS, 2007., 2007, Berlin, Heidelberg. **Proceedings...** Springer-Verlag, 2007. p.202–212. (ER'07).

SUBRAMANIAN, S. V.; KAWACHI, I. Income Inequality and Health: what have we learned so far? **Epidemiol Rev**, [S.l.], v.26, n.1, p.78–91, July 2004.

TANG, J.; MUSOLESI, M.; MASCOLO, C.; LATORA, V. Temporal distance metrics for social network analysis. In: ACM WORKSHOP ON ONLINE SOCIAL NETWORKS, 2., 2009, New York, NY, USA. **Proceedings...** ACM, 2009. p.31–36. (WOSN '09).

TEJADA, S.; KNOBLOCK, C. A.; MINTON, S. Learning object identification rules for information integration. **Inf. Syst.**, Oxford, UK, UK, v.26, n.8, p.607–633, 2001.

TERVEEN, L.; MCDONALD, D. W. Social matching: a framework and research agenda. **ACM Trans. Comput.-Hum. Interact.**, New York, NY, USA, v.12, n.3, p.401–434, 2005.

WANG, C.; HAN, J.; JIA, Y.; TANG, J.; ZHANG, D.; YU, Y.; GUO, J. Mining advisor-advisee relationships from research publication networks. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 16., 2010, New York, NY, USA. **Proceedings...** ACM, 2010. p.203–212. (KDD '10).

WASSERMAN, S.; FAUST, K. **Social Network Analysis**: methods and applications. [S.l.]: Cambridge University Press, 1994.

WATTS, D. J. **Six Degrees**: the science of a connected age. [S.l.]: W. W. Norton & Company, 2003.

WENG, S.-S.; CHANG, H.-L. Using ontology network analysis for research document recommendation. **Expert Syst. Appl.**, Tarrytown, NY, USA, v.34, n.3, p.1857–1869, 2008.

XIANG, L.; YUAN, Q.; ZHAO, S.; CHEN, L.; ZHANG, X.; YANG, Q.; SUN, J. Temporal recommendation on graphs via long- and short-term preference fusion. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 16., 2010, New York, NY, USA. **Proceedings...** ACM, 2010. p.723–732. (KDD '10).

XU, Y.; HAO, J.; LAU, R. Y.; MA, J.; XU, W.; ; ZHAO, D. A Personalized Researcher Recommendation Approach in Academic Contexts: combining social networks and semantic concepts analysis. In: PACIFIC ASIA CONFERENCE ON INFORMATION SYSTEMS (PACIS), 2010, Taipei, Taiwan. **Proceedings...** [S.l.: s.n.], 2010. (PACIS '10).

YAN, S.; LEE, D. Toward alternative measures for ranking venues: a case of database research community. In: ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES, 7., 2007, New York, NY, USA. **Proceedings...** ACM, 2007. p.235–244. (JCDL '07).

ZAIANE, O. R.; CHEN, J.; GOEBEL, R. DBconnect: mining research community on dblp data. In: WEBKDD AND 1ST SNA-KDD 2007 WORKSHOP ON WEB MINING AND SOCIAL NETWORK ANALYSIS, WEBKDD/SNA-KDD '07, 9., 2007, New York, NY, USA. **Proceedings...** ACM, 2007. p.74–81.

ZAÏANE, O. R.; CHEN, J.; GOEBEL, R. Mining Research Communities in Bibliographical Data. In: ZHANG, H.; SPILIOPOULOU, M.; MOBASHER, B.; GILES, C. L.; MCCALLUM, A.; NASRAOUI, O.; SRIVASTAVA, J.; YEN, J. (Ed.). **Advances in Web Mining and Web Usage Analysis**. Berlin, Heidelberg: Springer-Verlag, 2009. p.59–76.

ZANARDI, V.; CAPRA, L. Social ranking: uncovering relevant content using tag-based recommender systems. In: ACM CONFERENCE ON RECOMMENDER SYSTEMS, RECSYS '08, 2008., 2008, New York, NY, USA. **Proceedings...** ACM, 2008. p.51–58.