sbbd:05

Contribuição de Pesquisa entre Comunidades como Indicador de Influência

Thiago H. P. Silva, Lais M. A. Rocha
Mirella M. Moro, Ana Paula Couto da Silva

Departamento de Ciência da Computação, Universidade Federal de Minas Gerais
Belo Horizonte – MG – Brasil

{thps,laismota,mirella,ana.coutosilva}@dcc.ufmg.br

Resumo. *No contexto de extração de informações, este artigo propõe uma nova métrica de influência científica a partir de dados bibliográficos. Em particular, o objetivo é medir o grau de influência de pesquisadores através da avaliação de relações entre comunidades, que são formadas a partir dos eventos em que publicam. Parte-se do princípio que cada pesquisador tem uma comunidade-base em que ele/ela apresenta maior influência. Quando tal pesquisador trabalha em uma comunidade diferente (além da comunidade-base), leva novos conhecimentos a essa comunidade e há transferência de influência, aumentando a qualidade global das comunidades. Ao medir tal transferência, pretendemos medir a influência de pesquisadores nas respectivas comunidades.*

Abstract. *In the context of information extraction, this paper proposes a new influence metric derived from bibliographic data. Specifically, the goal is to measure the degree of influence of researchers by evaluating the links between communities, which are formed by their publication venues. Each researcher has a base community where he/she presents greater influence. Then, when such researcher works on a different community (besides the base community), he/she takes new knowledge to that community and transfers influence, which improves the global quality of the communities. By pondering such transfer, we measure the influence of researchers in their and across communities.*

1. Introdução

Uma tarefa central em bancos de dados é extrair informações relevantes a partir de seus dados. Nas últimas duas décadas, a comunidade científica tem tentado definir métricas para extrair conhecimento (geralmente na forma de rankings) a partir de dados bibliográficos, a chamada *bibliometria*. Métricas tradicionais incluem volume de artigos produzidos e seu número de citações; e uma das métricas mais utilizadas é o *h-index*, o qual combina as duas informações [Hirsch 2005]. Porém, todas essas métricas puramente quantitativas têm sofrido grandes questionamentos (e.g., [Hicks et al. 2015]).

Para contrapor essas métricas centradas no pesquisador [Bollen et al. 2009], uma solução é avaliar como o indivíduo se relaciona com seus pares e as comunidades definidas pelos mesmos. Especificamente, estudos recentes de redes sociais acadêmicas têm avaliado o desempenho de cada pesquisador individualmente e agrupado em times, comunidades, programas de pós-graduação entre outros [Freire and Figueiredo 2011, Lopes et al 2011, Newman 2004]. Outras análises exploram

aspectos qualitativos como a influência dos veículos de publicação [Garfield 1999], comunidades [Silva et al 2014, Silva et al. 2015], redes de colaborações [Brandão et al. 2014], perfis acadêmicos [Gonçalves et al 2014, Lima et al 2015], e outros.

Considerando a perspectiva social, alguns estudos focam em analisar como posições privilegiadas na rede são definidos de diferentes maneiras. Por exemplo: Granovetter (1973) introduziu a ideia de *weak ties*; Newman (2000) explorou o número de caminhos mais curtos entre pares de nós para medir a influência sobre o fluxo de informações entre os indivíduos; e Burt (2004) chamou de *brokers* as pessoas que constroem capital social ao se posicionarem em pontos estratégicos na rede (*structural holes*). De maneiras diferentes, tais estudos enfatizam a importância de construir pontes e de conectar nós distantes da rede social. Ou seja, em redes sociais acadêmicas, pesquisadores que conectam diferentes grupos devem trazer mais influência a esses grupos. A hipótese base é explorada em outros contextos (e.g., em dados econômicos), e nosso estudo mostra que esta também pode ser aplicada como indicador de influência na área de Computação.

Este trabalho contribui para esta discussão em dois aspectos: (*i*) mostra que é possível ampliar o conceito de construção de pontes ao explorar a relação pesquisador-comunidade e (*ii*) propõe uma nova estratégia que mensura a influência que, ao mesmo tempo, lida com o problema da discrepância entre comunidades ao projetar a influência sob uma mesma comunidade. Especificamente, analisamos a propagação de influência entre *comunidades*, que são formadas por pesquisadores que compartilham interesses comuns pela publicação de artigos na mesma conferência. Para isso, definimos que cada pesquisador possui uma *comunidade base* como sendo aquela em que ele possui o melhor desempenho. Então, seguimos os conceitos de diversidade e novidade [Burt 2004], e consideramos que quando um pesquisador trabalha em uma comunidade diferente (além da comunidade base), ele/ela transfere novos conhecimentos de seu contexto para outros domínios. Ou seja, há transferência de novas ideias, técnicas e metodologias que resultam em contribuições para a evolução da qualidade de pesquisa como um todo.

O novo indicador proposto, chamado **3c-index** (*Cross-Community Contribution*), utiliza uma função genérica de *score* para avaliar o grau de influência do pesquisador em diferentes comunidades, e depois projeta tal valor em sua comunidade base. As principais vantagens são: *flexibilidade*, pois o *3c-index* permite o uso de quaisquer funções de score (e.g., número de citações, volume, h-index, socialibilidade ou número de coautores, volume de produção recente e métricas de redes complexas); e *equidade*, pois releva os diferentes perfis das comunidades envolvidas através da projeção em uma única comunidade. Nossa análise também mostra que o *3c-index* traz resultados diferentes nos ranqueamentos em relação aos índices tradicionais e identifica quais são os pesquisadores que estão em buracos estruturais, i.e., aqueles que mais atuam em diferentes partes da rede tornando-a mais forte. A nossa validação experimental também reporta que os resultados do *3c-index* superam índices tradicionais na tarefa de ranquear pesquisadores com premiações dadas por suas contribuições e inovações em suas comunidades.

Finalmente, é importante notar que o 3c-index não deve ser utilizado sozinho. O seu objetivo principal é fornecer uma análise *complementar* cujo resultado não pode ser diretamente aferido a partir de uma simples consulta SQL aos dados bibliográficos. Segundo Hicks et al. (2015) , é necessário que seja utilizado por especialistas junto a outras avaliações qualitativas.

2. Trabalhos relacionados

É sempre difícil avaliar pessoas de acordo com algum critério. Neste trabalho focamos na relação social entre os pesquisadores e suas comunidades. Um dos primeiros trabalhos que lidam com a perspectiva social foi realizado por Granovetter (1973) através da ideia de *weak ties* como sendo as relações que unem diferentes partes da rede através da construção de pontes. Newman (2000) considerou o papel de um pesquisador na rede e explorou o número de caminhos mais curtos de um nó central para outros nós e, assim, mediu a influência sobre o fluxo de informações entre os indivíduos (níveis de *brokering*). De forma similar, Burt (2004) chama de *brokers* as pessoas que constroem capital social ao se posicionarem em pontos estratégicos na rede (*structural holes*). Nosso trabalho contribui para esta discussão ao analisar a influência propagada por pesquisadores entre comunidades, i.e., mostramos que podemos ampliar o conceito da construção de pontes ao explorar a relação pesquisador-comunidade.

Considerando a área de Computação, Freire e Figueiredo (2011) exploraram as colaborações *externas* de indivíduos e grupos. Eles identificaram os indivíduos e grupos influentes através da intensidade de seus relacionamentos com indivíduos de fora do grupo. De forma similar, Silva et al. (2014) exploraram o conceito de *comunidade* para ranquear veículos de acordo com o grau das relações externas de seus membros. Os resultados indicam que pesquisadores que publicam em outras comunidades possuem maior probabilidade de introduzir novas ideias de sua competência para outros contextos e, sendo assim, tais pesquisadores são considerados altamente influentes por ligarem diferentes partes da rede (ou seja, funcionam como *weak ties* ou *brokers* da rede). Em relação ao fluxo de contribuições entre comunidades, Silva et al. (2015) propuseram quatro métricas para mensurar a dinâmica entre pesquisadores em comunidades (*Permanency*, *Migration*, *Exclusivity* e *Plurality*). Aqui, propomos avaliar pesquisadores explorando o vínculo de um pesquisador com seu conjunto de comunidades e, a partir disso, explorar o grau de influência que pode ser transmitida entre diferentes comunidades.

Em termos de análise da produtividade, é necessário considerar os diferentes padrões de publicação existentes. Por exemplo, Gonçalves et al. (2014) caracterizaram a produtividade de pesquisadores e concluíram que há, realmente, perfis diferentes e bem definidos. Comunidades de pesquisadores também possuem padrões distintos quanto ao número de membros, áreas de pesquisa, taxas de publicações, etc. Para lidar com tal problema, Silva et al. (2015) estabeleceram propriedades (*Equality*, *Relativity* e *Temporality*) que devem ser discutidas na metodologia para uma comparação mais justa. Aqui, nós definimos a *comunidade base* do pesquisador como aquela onde ele possui o melhor desempenho, e a influência é medida na *perspectiva de cada pesquisador em relação à sua comunidade base, buscando uma avaliação relativamente mais justa*.

No contexto de ranqueamento, Silva et al. (2015) propuseram uma estratégia baseada em *grupos similares* formados por membros com características comuns em relação ao vínculo temporal com os veículos de publicação, argumentando que há mais justiça ao comparar perfis semelhantes. Lima et al. (2013) criaram uma estratégia genérica para pesquisadores pertencentes a múltiplas áreas através da projeção da produtividade sob uma mesma perspectiva. Similarmente, nós aplicamos uma estratégia de ranqueamento baseado em percentis para mapear os valores de diferentes comunidades para uma *comunidade base* do pesquisador. Além disso, corrigimos os valores ao definir o grau de

Tabela 1. Grau de influência entre comunidades de Christos Faloutsos.

Comunidade	Percentil	$infd_i$
SIGKDD	0.995	0
SIGMOD	0.92	0.07
SIGMETRICS	0.68	0.32
SIGAPP	0.48	0.52

influência existente entre as comunidades sob a perspectiva da contribuição de cada autor. Assim, quanto maior o grau de influência, maior será a probabilidade de que haja transferência de conhecimento do pesquisador para comunidades distintas.

3. Contribuição entre Comunidades

Pesquisadores podem ser agrupados em uma ou mais áreas de competência [Lima et al 2013, Lima et al 2015]. Então, a transferência de conhecimento entre áreas é fundamental para contribuir com a evolução da Ciência, pois possibilita aplicações de ideias conhecidas e amplamente utilizadas em um contexto para resolver problemas em outros domínios [Sun et al 2013]. Neste trabalho definimos a métrica **3c-index** que identifica a transferência de conhecimento entre diferentes comunidades. O novo índice mede a influência dos pesquisadores de acordo com as suas especialidades (definido como *grau de influência*) para outras comunidades.

Uma característica importante do 3c-index consiste em lidar com justiça ao comparar comunidades com diferentes padrões de publicação. Para isso, definimos o conceito de *comunidade base* de um pesquisador como aquela em que ele possui o melhor desempenho. A escolha da comunidade base é definida em termos de percentis, i.e., de acordo com as posições dos pesquisadores no ranqueamento. Formalmente, p_i^c é o percentil do pesquisador *i* na comunidade *c* definido por: $p_i^c = \frac{l_i^c + 0.5e_i^c}{N^c}$, onde N^c é o número de pesquisadores na comunidade c, l_i^c e e_i^c o número dos pesquisadores com valores no ranqueamento inferiores ou iguais ao do pesquisador i, respectivamente. Por exemplo, para uma comunidade com 100 pesquisadores com scores distintos (sem empates), o pesquisador na posição 10 possui $l_i^c = 89$ e $e_i^c = 1$, resultando no percentil de 89.5%. Após o cálculo de p_i^c para todas as comunidades em que o pesquisador publica, defini-se *comunidade base* b_i como sendo aquela onde o pesquisador possui o maior valor para o percentil, i.e., $b_i = \text{argmax}_{c \in C} p_i^c$.

Tendo definido a *comunidade base* de um pesquisador, mede-se então o grau de influência $infd_i$ como a diferença entre os percentis (i.e., $infd_i = b_i - p_i^c$). Por exemplo, considerando todos os autores que publicam em um evento promovido por um SIG[1] como uma comunidade científica, pode-se identificar a influência propagada por seus membros para as demais comunidades. Para ilustrar, a Tabela 1 mostra o grau de influência propagada pelo pesquisador *Christos Faloutsos*: sua comunidade base é a SIGKDD, no qual p_i^c apresenta maior valor; SIGMOD evidencia grau similar; e com graus moderados de contribuições estão SIGAPP e SIGMETRICS. Note que $infd_i = 0$ para a comunidade base pois, por definição, não há transferência de conhecimento neste caso. Assim, o grau de influência captura os relacionamentos importantes dos pesquisadores que atuam como *pontes* entre comunidades.

[1]Association for Computing Machinery Special Interest Groups. `http://www.acm.org/sigs`

Dado que comunidades científicas possuem perfis distintos (i.e., número de artigos, taxas de publicações e citações, etc), é necessário um fator de normalização. Conforme o modelo de ranqueamento proposto por Lima et al. [Lima et al 2013], nossa estratégia considera os valores dos percentis dos pesquisadores *projetados* para sua comunidade base. O *3c-index* de um pesquisador i é então definido como

$$3c{-}index(i) = score(b_i) + 2\sum_{c \in C}(b_i - p_i^c)score(f_b(p_i^c)),$$

onde $score(x)$ é o valor no ranqueamento na posição de percentil x, $f_b(x)$ é a função de projeção que mapeia o percentil x para o percentil correspondente no ranqueamento da comunidade base b. Dessa forma, o valor final consiste no valor obtido na sua comunidade base ($score(b_i)$) somado aos valores do ranqueamento provenientes de contribuições em comunidades externas. Note que a estratégia de ranqueamento é genérica e possibilita o uso de qualquer função de *score* como: número de citações, volume de publicações, h-index, número de alunos formados, número de coautores, métricas de análise de redes complexas, entre outros.

4. Metodologia

Esta seção apresenta a metodologia a ser utilizada na avaliação do nosso novo índice. Especificamente, o processo de avaliação é apresentado em duas perspectivas. Na primeira (Seção 5), endossamos a existência da troca de contribuição entre comunidades, investigamos as formações das comunidades em relação aos membros que a possuem como *comunidade base*, e avaliamos a independência entre a nossa proposta e as métricas padrão. Na segunda (Seção 6), validamos a nossa abordagem para ranquear pesquisadores relevantes (vencedores de *ACM Awards*) ao compararmos com as métricas padrão.

Para realizar tais avaliações, primeiro é necessário definir uma base de dados que contenha todas as informações necessárias sobre os pesquisadores a serem avaliados. Para avaliar e validar o índice proposto, também é necessário: definir métricas padrão para função de score e com as quais o desempenho do novo índice será comparado; definir uma tabela verdade para verificar se os resultados do índice fazem sentido; e definir métricas de avaliação considerando o ranqueamento gerado pelo índice. A seguir, define-se cada uma dessas partes.

Base de Publicações. O 3c-index é baseado no conceito de comunidade. Para fins de avaliação experimental, considera-se as comunidades definidas por um sub-conjunto de ACM SIGs. Cada SIG organiza um ou mais eventos científicos sobre tópicos de interesse específicos. Além de promoverem grandes eventos, essas SIGs concedem prêmios a seus membros devido a serviços prestados, contribuições e inovações (*ACM Awards*). Especificamente, são 12 conferências como comunidades, as quais são consideradas de alta qualidade na área de Ciência da Computação e também são alvos de pesquisadores de alta qualidade. A Tabela 2 apresenta as comunidades e suas estatísticas. Para cada evento, foram coletadas suas listas de artigos a partir da DBLP[2] considerando o intervalo de tempo [2001,2010]. Também foram coletadas as citações dos trabalhos na plataforma do *Google Scholar*[3], casando-se 4-tuplas formadas por {título, ano, autores, veículo}. Por fim, a base de dados contém mais de dezoito mil autores, mais de cem mil artigos com

[2]DBLP: `http://www.informatik.uni-trier.de/~ley/db`
[3]GS: `http://scholar.google.com`

Tabela 2. Estatísticas da base de dados compreendendo o intervalo [2001,2010].

SIG	Aut.	Art.	$\frac{\text{Aut.}}{\text{Art.}}$	Cit. (10^{-3})	$\frac{\text{Cit.}}{\text{Aut.}}$	$\frac{\text{Cit.}}{\text{Art.}}$	#Awards
SIGACCESS	849	418	2,03	7,9	9,25	18,79	2
SIGAda	169	135	1,25	0,7	4,21	5,27	15
SIGAPP	6732	3078	2,19	45,2	6,71	14,68	7
SIGCOMM	816	338	2,41	104,7	128,35	309,86	14
SIGCSE	1957	1143	1,71	23,1	11,78	20,17	25
SIGDOC	460	316	1,46	3,0	6,61	9,62	9
SIGIR	2313	1528	1,51	85,2	36,86	55,79	5
SIGKDD	2172	1075	2,02	109,3	50,33	101,69	17
SIGMETRICS	1053	449	2,35	31,2	29,63	69,49	5
SIGMOBILE	748	276	2,71	66,9	89,41	242,31	7
SIGMOD	2196	1098	2	103,9	47,32	94,65	29
SIGUCCS	838	655	1,28	1,7	2	2,56	5
Todas	18511	10509	1,76	582,8	31,49	55,46	137

mais de meio milhão de citações recebidas. As médias são de 1,8 de autores por artigo e de 55,5 citações por artigo.

Métricas padrão. São considerados três índices bibliométricos bem conhecidos: volume de publicações, número de citações e h-index. Tais índices são amplamentes usados em plataformas de busca orientados à pesquisa como *Google Scholar*, *Microsoft Academic Search* e *AMiner*. Como a estratégia de ranqueamento é genérica e possibilita o uso de qualquer função de *score* (Seção 3), a seguir, simplificamos a nomenclatura como 3c-citações, 3c-volume e 3c-h-index para indicar que estamos usando a métrica 3c-index com as funções de *score*, respectivamente, das métricas: número de citações recebidas, volume de publicações e h-index.

Tabela Verdade. Ranquear pesquisadores é um grande desafio devido à falta de um consenso de quais são as métricas ideais para tornar o processo de decisão justo. Alternativas de avaliação consistem em determinar grupos de referências contendo membros com relevância inquestionável em suas avaliações. Por exemplo, Hirsch (2005) usou os vencedores do prêmio Nobel para avaliar sua métrica (h-index), e Lima et al. (2013) usaram os níveis de bolsas governamentais concedidas. De forma similar, nós construímos um *ground-truth* de pesquisadores influentes como sendo os 137 (dos 18.511) pesquisadores vencedores de pelo menos um *ACM Award*. Nós assumimos que esses seletos pesquisadores possuem relevância e impacto inquestionáveis para as comunidades que os premiaram.

Métricas de Avaliação. Para avaliação, são considerados dois tipos de métrica: um para comparar o resultado do ranqueamento, e outro para validar o ranqueamento.

Com o intuito de avaliar o comportamento do 3c-index, consideramos a correlação de Spearman ρ (mede a dependência estatística entre dois rankings) e a distância de Jaccard (interseção sobre união) para mensurar a dissimilaridade entre dois conjuntos. O objetivo principal é verificar a independência entre a nossa abordagem e as métricas padrão, i.e., verificar se a proposta traz novidade com possibilidade de ser usada como forma completementar para uma análise mais completa.

Para melhor investigar a eficácia do 3c-index para ranqueamento, utilizamos a métrica DCG (*Discounted Cumulative Gain*). Esta métrica mede a qualidade do ranqueamento de acordo com uma escala graduada com um fator logarítmico de penalização para itens relevantes recuperados tardiamente [Järvelin and Kekäläinen 2002]. Formalmente,

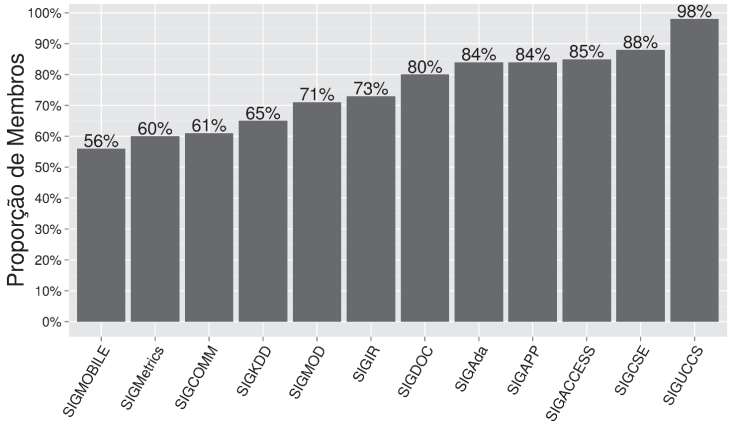


Figura 1. Porcentagem de autores (com mais de um artigo) que possuem a própria comunidade como sua *comunidade base* de acordo com o h-index.

o DCG na posição de ranqueamento k é definido como $DCG@k = g_1 + \sum_{i=2}^{k} \frac{g_i}{log_2(i)}$, onde g_i denota uma relevância binária (i.e., 1 se o pesquisador é vencedor de um *ACM Award*, 0 caso contrário). Nós usamos a versão normalizada (nDCG), que é obtida ao dividirmos o DCG@k pelo melhor possível ranqueamento no mesmo corte k.

5. Análise

Nesta seção investigamos a aplicabilidade do 3c-index para o contexto de comunidades científicas. A Seção 5.1 verifica a dependência das comunidades em relação à formação dos seus membros de acordo com o conceito de *comunidade base*, i.e., estamos interessados em endossar que há evidências o suficiente para aplicarmos a métrica em tais contextos. A Seção 5.2 mostra os pesquisadores mais influentes nos SIGs, bem como o reconhecimento deles pelas comunidades da ACM. Então, a Seção 5.3 compara a nossa proposta com índices tradicionais com objetivo de mostrar a independência entre eles.

5.1. Transferência de Conhecimento

A questão a ser respondida nesta seção é se existe transferência de conhecimento entre comunidades. Uma vez que parte do valor agregado da métrica proposta consiste em quantificar o grau de influência dos pesquisadores em comunidades, temos então que analisar como é composta cada comunidade em relação à participação de membros externos. Consideramos como função de score o percentil de acordo com o ordenamento pelo h-index[4] e que somente os membros com mais de uma publicação em uma conferência formam a comunidade da mesma. A Figura 1 mostra a proporção de participantes em cada SIG que possuem a própria comunidade como base. Por exemplo, 56% dos membros da SIGMOBILE a possuem como sua comunidade base, i.e., ela recebe contribuição externa de 44% dos seus autores. Em contrapartida, a SIGUCCS possui apenas 2% de contribuição externa, indicando que a conferência tende a possuir uma comunidade extremamente fechada (de fato, o foco desta SIG é dar suporte logístico aos serviços de tecnologia da informação para instituições de ensino[5]). As conferências restantes pos-

[4]Foram obtidos valores similares para o número de citações recebidas e volume.

[5]SIGUCCS: `http://www.siguccs.org/`

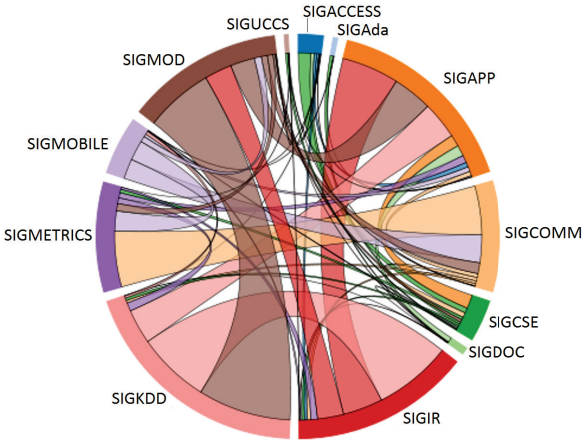


Figura 2. Transferência de Conhecimento entre Comunidades (ACM SIGs).

suem proporções de 12% a 40% de membros externos, ressaltando a existência da troca de conhecimento entre comunidades diferentes.

Para uma visualização mais clara da interação entre as comunidades, a Figura 2 mostra a transferência de influência entre as SIGs. Por exemplo, se a comunidade base de um pesquisador for SIGMOD e ele publicar no SIGKDD, então o seu valor do 3c-index (usando o h-index) é contabilizado como partindo do SIGMOD para o SIGKDD (i.e., nesse caso o valor acumulado somente é atribuído para a SIGKDD que foi o destino da contribuição). Sendo assim, cada parte da visualização circular é proporcional ao acúmulo de contribuições recebidas. As conferências SIGUCCS, SIGAda e SIGDOC possuem os menores valores de transferência de conhecimento, enquanto as conferências SIGKDD, SIGAPP, SIGIR e SIGMOD possuem os maiores valores. O aspecto do gráfico é bem coerente em relação às áreas que possuem uma maior proximidade dos seus programas como SIGMOD/SIGKDD, SIGMOD/SIGIR e SIGCOMM/SIGMETRICS. Finalmente, é possível notar que existe interações entre todas as comunidades.

5.2. Pesquisadores Influentes nos SIGs

Nesta seção verifica-se quais são os pesquisadores considerados mais influentes nos SIGs, confrontando essa lista com as distinções de reconhecimento pelos seus trabalhos de acordo com a ACM. A Tabela 3 mostra os top 10 pesquisadores de acordo com o 3c-index usando como funções de score o número de citações recebidas, volume e h-index. A notação é a seguinte: os pesquisadores que já ganharam premiações devido às suas contribuições e inovações para uma das comunidades (i.e., o prêmio é dado pelo próprio SIG) estão em negrito; e outros reconhecimentos da ACM (*fellow*, *distinguished* e *senior*)[6] estão marcados com símbolos.

Os resultados mostram pesquisadores conhecidos na área de Computação. Em especial, o 3c-volume e o 3c-h-index têm em suas primeiras colocações os pesquisadores agraciados com prêmios dados por suas comunidades (cinco primeiros para volume e quatro primeiros para o h-index). Este resultado é importante pois há apenas 137 de 18.511 pesquisadores de nossa base que ganharam tal premiação. Na próxima subseção compa-

[6]ACM Membership: `http://awards.acm.org/grades-of-membership.cfm`

Tabela 3. Pesquisadores melhor ranqueados de acordo com 3x-index nos SIGs.

Posição	3c-citações	3c-volume	3c-h-index
1º	**Scott Shenker**♯	**W. Bruce Croft**♯	**Jiawei Han**♯
2º	Ion Stoica♯	**Christos Faloutsos**♯	**Christos Faloutsos**♯
3º	M. Frans Kaashoek♯	**Surajit Chaudhuri**♯	**Scott Shenker**♯
4º	David R. Karger♯	**Jiawei Han**♯	**W. Bruce Croft**♯
5º	Sylvia Ratnasamy	**Scott Shenker**♯	ChengXiang Zhai‡
6º	Mark Handley	ChengXiang Zhai‡	**Surajit Chaudhuri**♯
7º	Paul Francis	Philip S. Yu♯	Wei-Ying Ma‡
8º	Richard M. Karp♯	Zheng Chen†	Zheng Chen†
9º	**Jon M. Kleinberg**♯	Divesh Srivastava♯	Philip S. Yu♯
10º	Dina Katabi♯	Leif Azzopardi	Divesh Srivastava♯

Em negrito os ganhadores do ACM Awards, ♯*ACM fellow*, ‡*ACM distinguished scientist*, †*ACM senior member*.

ramos o 3c-index em relação à composição de pesquisadores nas primeiras posições do ranqueamento (i.e., se são semelhantes), bem como os seus posicionamentos.

5.3. Comparação com Métricas Padrão

Outra questão a ser respondida é se existe independência entre o índice proposto e as métricas padrão. Esta questão é importante para mostrar que a nossa estratégia captura características diferentes e, dessa forma, pode ser usada de forma *complementar* com outros índices. Primeiro consideramos o posicionamento dos pesquisadores de acordo com cada métrica. Para isso, usamos o coeficiente de Spearman que mede a intensidade da ordem (ao invés do valor) entre dois ranqueamentos. Em seguida, exploramos a formação do ranqueamento de cada métrica em relação à presença dos mesmos pesquisadores, i.e., verificamos se há dissimilaridade entre os conjuntos de pesquisadores que compõem as primeiras posições. Tal dissimilaridade é medida por 1 menos a divisão entre a interseção e união de dois conjuntos (Jaccard).

A Tabela 4 mostra as correlações de Spearman entre índices tradicionais e seus correspondentes na versão proposta 3c-index. Aqui, considera-se a correlação forte e muito forte para $\rho \geq 0.7$, moderada para $0.4 \leq \rho < 0.7$ e fraca para $\rho < 0.4$. Em relação à métrica do número de citações recebidas e o 3c-index (C e $3c$), não há correlação para o topo do ranqueamento até a décima posição[7], valores moderados na comparação até a quadragésima posição, e valores fracos para as demais. As correlações do volume (V e $3c$) e do h-index (H e $3c$) em relação aos seus correspondentes na métrica 3c-index são semelhantes, com valores moderados para o topo do ranking (posições até 10 e 20) e valores fracos para as demais. De acordo com tal posicionamento dos pesquisadores em cada ranqueamento, há um grau de *independência* entre as métricas mostrada pelas correlações fracas e moderadas.

Para investigar se a formação do ranqueamento é similar em relação à presença dos mesmos pesquisadores, a Tabela 5 mostra as distâncias de Jaccard entre a métrica proposta e os índices tradicionais. Exceto em relação às citações ($d_J(C, 3c)$) para 20 pesquisadores e para o h-index ($d_J(H, 3c)$) para 10 pesquisadores, todas as configurações possuem dissimilaridade de pelo menos 24%. Desta forma, as métricas também são independentes em relação à formação dos conjuntos de pesquisadores nas primeiras posições.

[7]Os 10 primeiros na métrica de citações estão posicionados entre os 16 primeiros pelo 3c-index, sendo somente dois pesquisadores com as mesmas posições nos dois ranqueamentos.

Tabela 4. Correlações de Spearman entre citações (C), volume (V), h-index (H) e seus 3c-index ($3c$). Ranqueamentos ordenados pela métrica original.

Posições	C e $3c$	V e $3c$	H e $3c$
10	0,04	0,57	0,67
20	0,56	0,67	0,66
40	0,45	0,27	0,12
60	0,27	0,13	0,06
80	0,30	0,09	0,13
100	0,34	0,21	0,16

Tabela 5. Distância de Jaccard para os conjuntos com 10, 20, 50 e 100 pesquisadores entre as métricas citações (C), volume (V), h-index (H), e seus 3c-index ($3c$).

#Pesq.	$d_J(C, 3c)$	$d_J(V, 3c)$	$d_J(H, 3c)$
10	0,33	0,46	0,18
20	0,10	0,26	0,40
40	0,33	0,40	0,43
60	0,29	0,48	0,38
80	0,24	0,43	0,40
100	0,28	0,35	0,40

6. Validação Experimental

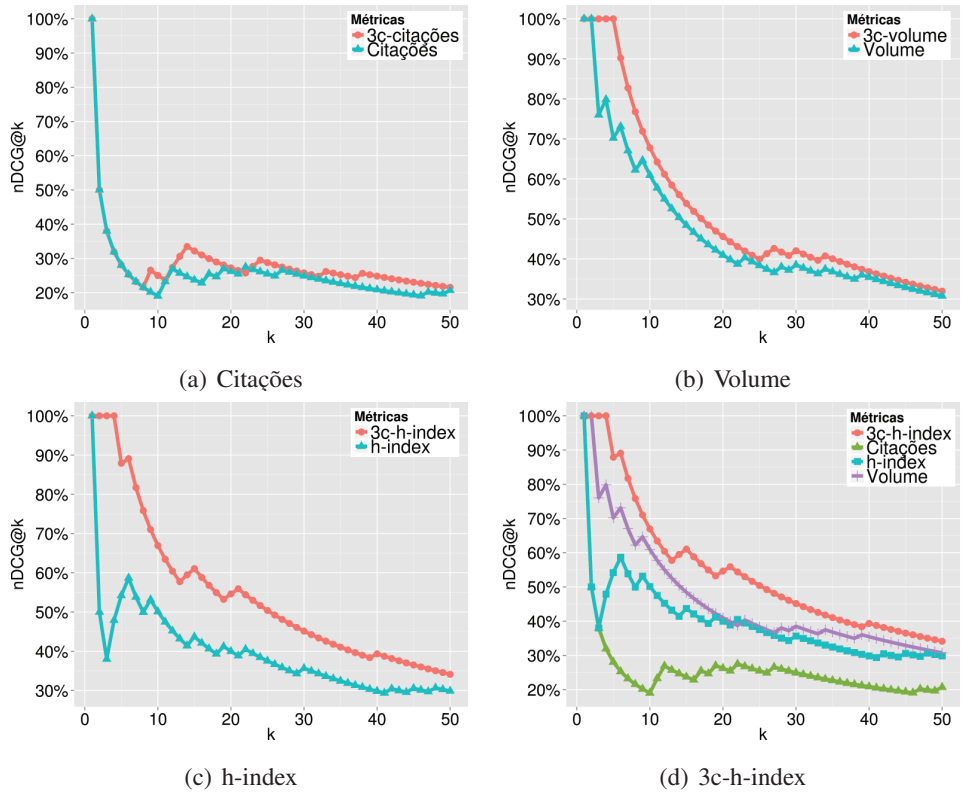
Após compararmos o comportamento do ranqueamento do 3c-index com as *métricas padrão*, nesta seção validamos o uso da métrica para ranquear pesquisadores considerados relevantes. Especificamente, a tarefa consiste em ranquear 18.511 pesquisadores com o objetivo de que as primeiras posições sejam ocupadas pelos 137 pesquisadores ganhadores de pelo menos um *ACM Awards*.

A Figura 3 compara o ranqueamento produzido pelo 3c-index com o número de citações, volume de publicações e h-index. Todos os gráficos consideram o valor *nDCG@k* obtido a cada corte *k*, variando da 1ª a 50ª posição (i.e., $k = 10$ mostra a comparação das métricas para ranquear somente os 10 primeiros pesquisadores). A figura mostra que o 3c-index supera consistentemente os índices tradicionais na maioria dos casos. Especificamente, considerando o número de citações recebidas (Figura 3(a)), há um empate entre as métricas nas primeiras posições até a 8ª posição e, então, o 3c-index supera ou é igual ao seu correspondente (exceto para a 22ª posição). Em relação ao volume de publicações (Figura 3(b)), há um empate entre as métricas nas primeiras duas posições e, a partir desta posição, a nossa proposta supera o seu correspondente.

É importante notar que as métricas baseadas no acúmulo de citações tendem a ser sensíveis a *outliers*, já que um só trabalho pode atrair muitas citações e, desta forma, podem existir ruídos nas avaliações que consideram conjuntos de artigos. O número de artigos publicados também é sensível quando, por exemplo, um autor pode assumir as primeiras posições ao ter um vasto volume em veículos específicos (i.e., maior taxa de aceitação e/ou menor qualidade). Como estamos considerando veículos de alto nível, o viés tende a ser diminuído.

Para contornar tais problemas, o h-index tenta controlar a sensibilidade ao considerar o volume e o número de citações recebidas ao mesmo tempo. De fato, o melhor desempenho comparativo da nossa proposta consiste na versão dada pelo h-index (Figura 3(c)), onde o 3c-index tem o melhor desempenho ao deixar nas primeiras quatro posições apenas pesquisadores relevantes e, desta forma, tende a superar o seu correspondente até o fim. A Figura 3(d) mostra o comparativo entre 3c-index (com o h-index como sua função de score) e as três métricas padrão resumindo o bom desempenho da nossa estratégia.

Os resultados inferiores das métricas tradicionais indicam que elas ranqueiam os pesquisadores premiados tardiamente (em posições mais afastadas do topo). Portanto, medir o impacto da transferência de conhecimento entre comunidades e lidar com os padrões de publicação de comunidades distintas é uma boa estratégia para ranquear pesquisadores influentes em comunidades.



(a) Citações (b) Volume

(c) h-index (d) 3c-h-index

Figura 3. Comparação entre 3c-index e os *baselines* de acordo com nDCG.

7. Conclusão

Neste trabalho apresentamos um novo índice de influência baseado na perspectiva social dos pesquisadores chamado 3c-index, que explora o grau de influência que pesquisadores exercem sobre comunidades. A proposta é robusta para lidar com o problema de ranqueamento justo de membros de comunidades com padrões distintos, bem como flexível no uso de qualquer função de score (seja baseada em volume, citações, métricas de redes complexas, etc). A análise experimental mostrou independência do 3c-index das métricas padrões em relação ao ranqueamento e formação do ranqueamento nas primeiras posições. Portanto, tais resultados endossam o uso da métrica de forma *complementar* para uma análise mais completa. Por fim, a nova abordagem teve o melhor desempenho para resolver um problema real de ranqueamento dos pesquisadores mais influentes (reconhecimento explícito por inovações e contribuições) em comunidades reconhecidas na área de Computação.

Como trabalho futuro, pretendemos aplicar o novo índice para investigar a influência entre as comissões especiais da SBC. Porém, existe a dificuldade extra de encontrar todas as informações com relação aos prêmios distribuídos pelas mesmas. O índice proposto também pode ser usado para identificar quais são os principais pesquisadores brasileiros na área de Computação ao se posicionarem bem entre instituições [Brandão and Moro 2012]. Outro estudo a ser realizado é a verificação do potencial de influência para utilização em sistemas de recomendação.

Agradecimentos. Trabalho parcialmente financiado por CNPq e FAPEMIG, Brasil.

Referências

Bollen, J., Van de Sompel, H., Hagberg, A., and Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS ONE*, 4(6):e6022.

Brandão, M. A. and Moro, M. M. (2012). Recomendação de colaboração em redes sociais acadêmicas baseada na afiliação dos pesquisadores. In *SBBD*, pages 73–80.

Brandão, M. A., Moro, M. M., and Almeida, J. M. (2014). Experimental evaluation of academic collaboration recommendation using factorial design. *JIDM*, 5(1):52–63.

Burt, R. S. (2004). Structural Holes and Good Ideas. *American Journal of Sociology*, 110(2):349–399.

Freire, V. P. and Figueiredo, D. R. (2011). Ranking in collaboration networks using a group based metric. *J. Braz. Comp. Soc.*, 17(4):255–266.

Garfield, E. (1999). Journal impact factor: a brief review. *Canadian Medical Association Journal*, 161(8):979–980.

Gonçalves et al, G. D. (2014). Characterizing scholar popularity: A case study in the Computer Science research community. In *JCDL*, pages 57–66, London, UK.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., and Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520(7548):429–431.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *PNAS*, 102(46):16569–16572.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Lima et al, H. (2013). Aggregating Productivity Indices for Ranking Researchers Across Multiple Areas. In *JCDL*, pages 97–106, Indianapolis, USA.

Lima et al, H. (2015). Assessing the profile of top brazilian computer science researchers. *Scientometrics*, 103(3):879–896.

Lopes et al, G. R. (2011). Ranking Strategy for Graduate Programs Evaluation. In *ICITA*, pages 253–260, Sydney, Australia.

Newman, M. E. (2004). Who Is the Best Connected Scientist? A Study of Scientific Coauthorship Networks. 650:337–370.

Silva, T. H. P., Moro, M. M., and Silva, A. P. C. (2015). tc-index: a New Research Productivity index based on Evolving Communities. In *TPDL*, Poznań, Poland.

Silva et al, T. H. P. (2014). Community-based Endogamy as an Influence Indicator. In *JCDL*, pages 67–76, London, UK.

Silva et al, T. H. P. (2015). Authorship Contribution Dynamics on Publication Venues in Computer Science: an Aggregated Quality Analysis. In *SAC*, Salamanca, Spain.

Sun et al, X. (2013). Social dynamics of science. *Sci. Rep.*, 3(1069).