Influência da ambiguidade de nomes na centralidade de redes de coautoria

Name ambiguity influences centrality of co-authorship networks

Rafael Garcia BARBASTEFANO[1]
Cristina SOUZA[1]
Juliana Maria de Sousa COSTA[1]
Patrícia Mattos TEIXEIRA[1]

Resumo

A ambiguidade de nomes de autores pode gerar uma série de problemas, tanto em estudos bibliométricos, quanto em Análise de Redes Sociais envolvendo esse tipo de dado. Trata-se de um problema ainda não resolvido, que vem se tornando cada vez mais complexo diante do aumento da produção científica no mundo, principalmente em países orientais, onde os nomes dos autores apresentam muitas similaridades, dificultando a correta identificação. O objetivo deste artigo é mostrar como o problema da ambiguidade de nomes influencia as métricas de centralidade utilizadas em Análise de Redes Sociais, que se destinam a identificar o papel desempenhado por um ator dentro de uma rede social. Este estudo baseou-se nas relações de coautoria de cerca de quase 30 mil artigos indexados na *Web of Science*. Para a análise, foram construídas três redes de coautoria fazendo uso de formas distintas de registro do nome dos autores. Os resultados apontaram diferenças significativas entre as redes, considerando-se as medidas de centralidade de grau (*degree*), proximidade (*closeness*) e intermediação (*betweenness*). O trabalho mostra o quanto é importante uma meticulosa etapa de normalização de nomes e chama atenção para a necessidade de maior discussão da questão no ambiente da academia, em busca de alternativas de solução.

Palavras-chave: Ambiguidade de nomes. Análise de redes sociais. Coautoria. Colaboração científica.

Abstract

Name ambiguity can create several problems in bibliometrics and Social Network Analysis studies related to this type of data. It is an unsolved issue that has become more complex in recent years due to increasing worldwide scientific publication, more specifically in far eastern countries. The aim of the article was to show how name ambiguity influences centrality metrics in Social Network Analysis. The study was based on co-authorship relations of approximately 30 000 articles indexed in Thomson Reuters/Web of Science. Three co-authorship networks were developed using different ways to record the authors' names. The results showed significant differences among the networks when considering degree centrality, closeness and betweenness. The study points out the importance of standardizing names and the need to further discuss the issue in the academic environment in an endeavor to find alternative solutions.

Keywords: *Name ambiguity. Social network analysis. Co-authorship. Scientific collaboration.*

[1] Ministério da Educação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Programa de Pós-Graduação em Engenharia de Produção e Sistemas. Av. Maracanã, 229, Bloco E, 5º andar, 20271-110, Rio de Janeiro, RJ, Brasil. Correspondência para/*Correspondence to*: R.G. BARBASTEFANO. *E-mail*: <rgb@cefet-rj.br>.

Recebido em 18/3/2014, reapresentado em 30/10/2014 e aceito para publicação em 18/11/2014.

Introdução

A colaboração científica tem sido objeto de diversos estudos ao redor do mundo, abrangendo diferentes campos do conhecimento (Barabasi *et al*., 2002; Glanzel, 2002; Li-Chun *et al*., 2006; Palla *et al*., 2007; Jiang, 2008). Conforme apontado por Çavusoglu e Türker (2013), trata-se de temática que tem recebido cada vez mais atenção, uma vez que indicadores de colaboração nacional e internacional têm sido utilizados para avaliar as universidades e seus pesquisadores, bem como, para tentar melhor compreender os efeitos dessas interações sobre os resultados das pesquisas realizadas. De acordo com Abbasi *et al*. (2010), trata-se do indicador mais visível e acessível para identificação das relações de colaboração no ambiente da academia.

De modo geral, os estudos sobre coautoria têm sido feitos sob duas abordagens: a primeira buscando compreender as razões que levam os autores a colaborarem entre si e as consequências dessas interações; e a segunda, baseando-se na ideia de que a coautoria forma uma rede social de pesquisadores (Acedo *et al*., 2006). A representação dessa rede de pesquisadores pode ser feita através de um grafo em que os vértices (nós) indicam os diversos autores, enquanto as arestas (ligações entre os nós) representam as publicações realizadas em conjunto. Trata-se, nesse caso, de rede de um "modo", ou seja, formada por uma única classe de vértices, que são os autores dos artigos publicados (Tomaél & Marteleto, 2013).

A partir da estrutura das redes de coautoria é possível identificar informações relacionadas à organização e dinâmica da pesquisa em uma determinada área do saber, à evolução das relações de colaboração, ao padrão de relação entre os autores e ao papel desempenhado por cada pesquisador através de diversas métricas de centralidade utilizadas em Análises de Redes Sociais (ARS) (Newman, 2001; De Souza *et al*., 2012; Yan & Guns, 2014).

Os resultados obtidos com a aplicação de métricas de ARS em redes de coautoria, entretanto, são fortemente afetados pela correta identificação dos autores e de suas publicações em parceria. Abreviações, mudanças de nome, homônimos, grafias diferenciadas e registros incompletos ou equivocados são exemplos de fatores que podem levar à ocorrência de ambiguidades na estruturação de uma rede, demandando tratamento prévio de normalização dos nomes dos autores (Kang *et al*., 2009; Smalheiser & Torvik, 2009; Tang & Walsh, 2010).

Os seguintes tipos de erros podem ocorrer em função dessas ambiguidades (Wang *et al*., 2012): (a) falsos vértices negativos ou positivos: ausência ou inserção indevida de nós na rede; (b) falsas arestas negativas ou positivas: ausência ou inserção indevida de arestas na rede; e (c) vértices falsamente agregados ou desagregados: quando dois ou mais nós são considerados como um nó único na rede ou quando são indevidamente considerados separados.

A literatura tem abordado a variação de medidas de ARS em função de erros ou falhas nos dados utilizados para a configuração de redes sociais, porém sem estarem diretamente associados ao problema da ambiguidade de nomes e sem aplicação em redes de coautoria. Tais abordagens se relacionam com a robustez das métricas obtidas através de sub-redes geradas a partir de amostras, bem como erros de outras origens.

Borgatti *et al*. (2006), por exemplo, avaliaram como se comportavam as medidas de centralidade a partir da remoção e adição aleatória de vértices e/ou arestas em redes randômicas geradas pelo método de Erdös e Renyi, com tamanhos variando de 10 a 100 vértices. Costenbader e Valente (2003) e Smith e Moody (2013) investigaram o efeito decorrente da perda de dados na formação das redes e seu impacto sobre métricas de centralidade. Kossinets (2006), por sua vez, abordou ainda as diferenças que ocorrem nas medidas topológicas das redes.

Tratando especificamente do impacto da grafia dos nomes sobre propriedades de redes de coautoria, pode ser citado o trabalho de Barbastefano *et al*. (2013), que indicou grandes variações em densidade, grau médio; abrangência da componente gigante; distribuição de graus; distância média e diâmetro e coeficiente de clusterização de Watts-Strogatz, ao se considerarem os nomes dos autores a partir de diferentes métodos.

Mas como esse problema dos nomes interfere especificamente nas medidas de centralidade que mostram a importância do papel de cada autor dentro

http://dx.doi.org/10.1590/0103-37862015000300001

de determinada rede de coautoria? Diante dessa questão, o objetivo do presente artigo é apresentar as variações nas métricas de centralidade de grau (*degree*), de proximidade (*closeness*) e de intermediação (*betweenness*), em função da adoção de diferentes grafias dos nomes dos autores (nome completo, nome abreviado e sobrenome acompanhado da primeira inicial). O estudo foi baseado em dados de coautoria de um conjunto de 28 916 artigos sobre o tema sustentabilidade.

Espera-se, com esse trabalho, sensibilizar os pesquisadores que atuam com ARS sobre a necessidade de realização de uma cuidadosa e sistemática etapa de normalização de nomes, bem como alertar os leitores e tomadores de decisão sobre a importância de se conhecerem os métodos e procedimentos adotados nos diversos estudos realizados.

Procedimentos metodológicos

Este trabalho fez uso do método bibliométrico, a partir de informações extraídas de artigos recuperados na *Web of Science* (WoS) sobre o tema sustentabilidade. Além de sua relevância como base científica, abrangendo publicações oriundas dos mais diversos países, a WoS apresenta interface que permite uma rápida e prática importação dos dados, sendo amplamente utilizada em estudos da área de ciência da informação. Por sua vez, sustentabilidade é uma temática que tem despertado grande interesse, atraindo pesquisadores de diferentes nacionalidades e áreas do conhecimento, o que foi considerado apropriado em função do objetivo do estudo.

Para fazer a comparação das métricas de centralidade, foram geradas três redes de coautoria, baseadas no mesmo universo de artigos, utilizando os nomes dos autores de formas distintas, sem tratamento dos dados para evitar ambiguidades. A primeira rede foi construída a partir do nome completo dos autores conforme constante na WoS (Rede I). A segunda usou o nome abreviado também fornecido pela WoS (Rede II). A partir do nome abreviado, foi elaborada uma lista com o sobrenome acompanhado da primeira inicial de cada autor, obtendo-se assim a terceira rede (Rede III).

As redes foram construídas com base nas relações de coautoria de um total de 28 916 documentos classificados como artigos. Outros tipos de publicação não foram considerados. A obtenção desses documentos foi feita usando o termo *sustainability*. Apesar de o levantamento ter sido realizado em 2013, o limite temporal restringiu-se aos artigos publicados até o ano de 2012.

Para a identificação das informações necessárias para as redes de coautoria, foi necessária a construção prévia de redes de autoria, baseadas nos pares "Autor-Artigo", nos três formatos adotados (nome completo, abreviado e sobrenome acompanhado da primeira inicial). As redes de autoria constituem redes bipartidas, em que cada vértice é um autor ou um artigo. Por sua vez, os artigos são ligados aos seus autores por arestas. A partir da conversão dessas redes bipartidas, foram construídas as redes de coautoria.

Cada rede de coautoria gerou um Grafo G (V,E) no qual V significa o conjunto de vértices (no caso, autores dos artigos) e E o conjunto de arestas que representam as relações sociais (no caso, a publicação de um artigo em comum). Tanto as redes de autoria como as de coautoria foram feitas com o uso do *Software Pajek*, específico para análise de redes sociais (Nooy *et al*., 2005). O programa é gratuito e pode ser obtido em <http://pajek.imfm.si>.

No escopo deste estudo foram utilizadas as seguintes medidas de centralidade, que buscam identificar a importância relativa de um ator dentro da rede: centralidade de grau (*degree*); centralidade de proximidade (*closeness*); e centralidade de intermediação (*betweenness*). Uma breve descrição de cada métrica encontra-se a seguir. Informações sobre ARS podem ser obtidas em Wasserman e Faust (1994).

- Centralidade de Grau (*degree*): o grau de um vértice *v* corresponde ao número de arestas incidentes ou ao número de vértices adjacentes a ele, aqui denotado por *d(v)*. Em uma rede de coautoria esse grau vai indicar o total de atores que publicaram em parceria com determinado ator.

- Centralidade de Proximidade (*closeness*): a centralidade de proximidade é definida pela soma das distâncias geodésicas entre determinado vértice e todos os outros vértices do grafo. Essa medida indica a proximidade de determinado ator em relação aos demais atores da rede. Matematicamente, a centralidade de

proximidade pode ser definida por meio da seguinte equação (Nooy *et al*., 2005):

$$C_C(v) = \frac{n-1}{\sum_{t \in V \setminus v} d_G(v,t)}$$

onde $d_G(v,t)$ corresponde à distância geodésica entre os vértices v e t.

- Centralidade de intermediação (*betweenness*): A centralidade de intermediação atribui importância a um ator em função da passagem de fluxo por ele, para interligar outros dois atores da rede, através do menor caminho possível. Essa medida foi definida como (Freeman, 1979):

$$Bet(v) = \sum_{i \neq j \neq v, i \neq v} \frac{\sigma_{i,j}(v)}{\sigma_{i,j}}$$

onde $\sigma_{i,j}$ é o número de caminhos geodésicos que ligam os vértices *i* e *j*, enquanto $\sigma_{i,j}(v)$ representa o número de caminhos geodésicos que ligam os vértices *i* e *j*, passando por *v*.

Como cada uma dessas três medidas visa avaliar um aspecto específico, o posicionamento de um autor no *ranking* de importância dentro da rede pode variar conforme a métrica de centralidade adotada. Assim sendo, a análise comparativa foi feita a partir dos valores e do ordenamento dos vértices mais centrais (no caso, autores) dentro de cada uma das redes geradas.

Os estudos de verificação de robustez da centralidade fazem uso de correlações para comparar os dados de redes com erro e redes originais (Costenbader & Valente, 2003; Borgatti, 2006). Para comparar os coeficientes de centralidade das três redes, utilizou-se a correlação de Spearman, em abordagem semelhante à utilizada por Wang (2012) para comparar diferentes redes com erros. A correlação foi aplicada em três vetores relacionados às três redes, estabelecendo-se a correspondência entre o nome completo, nome abreviado e o sobrenome acompanhado da primeira inicial.

Resultados

A Tabela 1 mostra os resultados dos dez autores que apresentaram os maiores valores para as medidas de centralidade de grau, *closeness* e *betweenness* nas três redes analisadas. Nota-se que, considerando-se as três medidas de centralidade, alguns dos autores mais centrais na Rede I (nome completo) também aparecem como mais centrais na Rede II (nome abreviado). Já na Rede III (sobrenome acompanhado da primeira inicial), os autores mais centrais são diferentes, com exceção de "Folke, Carl" (único que aparece no *ranking* das medidas de centralidade nas três redes). Os outros nove autores da Rede III aparentam ter nomes asiáticos. Esse fato pode ser decorrente do aumento do número de publicações de países como a China, por exemplo, que apresentam grande número de pessoas com sobrenomes muito parecidos ou iguais. Dessa forma, quando a identificação dos nomes é feita apenas pelo sobrenome acompanhado da primeira inicial, diversos autores são agrupados como um único, aumentando os valores das medidas de centralidade desses vértices.

Comparando-se os graus dos autores que aparecem nas Redes I e II, percebe-se um aumento nos valores. Esse aumento é devido à junção de autores com sobrenomes iguais e iniciais semelhantes, de modo que, quando ocorreu a abreviação, os graus foram somados. Já quando se compara o resultado de "Folke, C" nas Redes II e III, o valor do grau diminui, provavelmente porque os autores com o mesmo sobrenome e mesma inicial possuíam vértices em comum, e/ou outros vértices em que estavam ligados também foram agrupados em um único vértice. Assim, o número de ligações de "Folke, C" com outros autores foi reduzido.

Outra observação importante é que na Rede III os graus dos autores apresentam valores bem mais elevados do que nas Redes I e II, com exceção de "Folke, C" conforme analisado anteriormente. A explicação para esse fenômeno decorre da contração dos nomes, que faz com que diferentes autores, que possuem o mesmo sobrenome acompanhado da mesma inicial, sejam agrupados como sendo um único autor. Um exemplo é o caso de "Li, X" que, como ilustrado na Figura 1, decorre da contração do nome completo e do nome abreviado de diferentes autores de acordo com a grafia indexada na *Thomson Reuters/Web of Science*. Assim, o autor "Li, X" da Rede III resulta do agrupamento de, respectivamente, 30 e 15 autores diferentes considerados nas Redes I e II. Com essa contração, as relações de coautoria de cada

http://dx.doi.org/10.1590/0103-37862015000300001

Tabela 1. Maiores valores de centralidade em cada uma das redes.

Grau					
Rede I		Rede II		Rede III	
Autor	Valor	Autor	Valor	Autor	Valor
Folke, Carl	152	Folke, C.	191	Li, Y	201
Steffen, Will	103	Lal, R.	125	Folke, C	188
Smith, Pete	102	Steffen, W.	125	Wang, Y	187
Lal, Rattan	100	Smith, P.	122	Wang, X	175
Tomich, Thomas P.	97	Chapin, F.S.	114	Li, X	173
Chapin, F. Stuart, III	91	Tomich, T.P.	102	Zhang, J	173
Olsson, Per	84	Twomlow, S.	95	Wang, J	167
Polasky, Stephen	84	van der Leeuw, S.	94	Liu, J	148
Rockstrom, Johan	83	Campbell, H.	91	Liu, Y	146
Olsson, Lennart	80	Olsson, L.	89	Zhang, L	138
Closeness					
Rede I		Rede II		Rede III	
Autor	Valor	Autor	Valor	Autor	Valor
Folke, Carl	0,0141	Folke, C.	0,0768	Li, X.	0,1448
Liverman, Diana	0,0135	Steffen, W.	0,0737	Liu, Y.	0,1416
Steffen, Will	0,0135	Wilson, J.	0,0735	Folke, C.	0,1415
Olsson, Per	0,0134	Olsson, P.	0,0732	Liu, J.	0,1409
Olsson, Lennart	0,0132	Walker, B.	0,0731	Wang, X.	0,1405
Carpenter, Stephen R.	0,0131	Leemans, R.	0,0731	Li, Y.	0,1405
Costanza, Robert	0,0131	Svedin, U.	0,0722	Wang, Y.	0,1402
Chapin, F. Stuart, III	0,0130	Tomich, T.P.	0,0720	Zhang, J.	0,1385
DeFries, Ruth	0,0130	Carpenter, S.R.	0,0719	Chen, J.	0,1384
Lebel, Louis	0,0130	Liverman, D.	0,0719	Wang, J.	0,1383
Betweenness					
Rede I		Rede II		Rede III	
Autor	Valor	Autor	Valor	Autor	Valor
Folke, Carl	0,00120	Folke, C.	0,01380	Li, X.	0,01535
Leemans, Rik	0,00085	Wilson, J.	0,00870	Folke, C.	0,01050
Worm, Boris	0,00079	Lal, R.	0,00717	Zhang, J.	0,01020
Lebel, Louis	0,00077	Olsson, L.	0,00614	Wang, Y.	0,01012
Olsson, Lennart	0,00069	Steffen, W.	0,00591	Liu, Y.	0,00961
Tomich, Thomas P.	0,00059	Leemans, R.	0,00554	Wang, J.	0,00947
Potting, Jose	0,00056	Ladha, J.K.	0,00537	Li, Y.	0,00937
Bai, Xuemei	0,00054	Wackernagel, M.	0,00530	Lal, R.	0,00930
Raven, Rob	0,00052	Zhang, Y.	0,00527	Wang, X.	0,00881
Lal, Rattan	0,00051	Haberl, H.	0,00527	Liu, J.	0,00825

Fonte: Elaborada pelos autores (2014).

Notas: Rede I: Nome Completo; Rede II: Nome abreviado; Rede III: Sobrenome acompanhado da primeira inicial.

um desses diferentes autores passam a ser agrupadas em torno de um único vértice, no caso "Li, X", fazendo com que as medidas de grau da Rede III se tornem mais elevadas.

Analisando-se os resultados do *closeness* para as três redes (Tabela 1), pode-se observar que o valor dessa medida para o autor "Folke, C" aumentou da Rede I para a II, e da Rede II para a III. De forma similar ao que aconteceu com a medida de grau, os valores do *closeness* da Rede III também são maiores do que os observados nas Redes I e II, considerando-se o *ranking* dos dez principais autores de cada rede. Como o *closeness* mede quanto um vértice está próximo dos demais, o agrupamento de autores devido à abreviação dos nomes provavelmente gerou um menor número de vértices intermediários, reduzindo também as distâncias entre os pares de

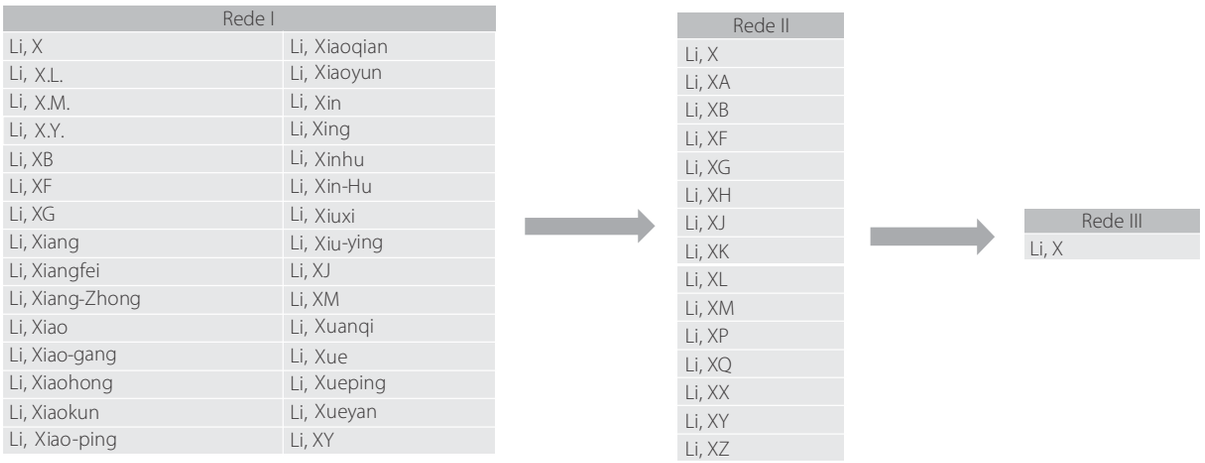

Figura 1. Como o nome "Li, X" se apresenta nas três redes.

Fonte: Elaborada pelos autores (2014).

Notas: Rede I: Nome Completo; Rede II: Nome abreviado; Rede III: Sobrenome acompanhado da primeira inicial.

Tabela 2. Medidas de centralidade de "Li, X" nas três redes de coautoria.

Rede I	*Close.*	*Bet.*	Grau	Rede II	*Close.*	*Bet.*	Grau	Rede III	*Close.*	*Bet.*	Grau
Li, X.	0,002393	0,000000	5	Li, X.	0,061379	0,001927	31	Li, X.	0,144841	0,015351	173
Li, X.L.	0,000082	0,000000	5	Li, X.A.	0,050923	0,000000	04				
Li, X.M.	0,000069	0,000000	4	Li, X.B.	0,069616	0,000423	28				
Li, X.Y.	0,000082	0,000000	5	Li, X.F.	0,056303	0,000355	15				
Li, X.B.	0,003345	0,000003	28	Li, X.G.	0,050783	0,000087	10				
Li, X.F.	0,000127	0,000000	4	Li, X.H.	0,053644	0,000187	16				
Li, X.G.	0,000093	0,000000	5	Li, X.J.	0,000063	0,000000	03				
Li, Xiang	0,000178	0,000000	6	Li, X.K.	0,054235	0,000000	04				
Li, Xiangfei	0,000082	0,000000	5	Li, X.L.	0,000094	0,000000	05				
Li, Xiang-Zhong	0,006122	0,000000	7	Li, X.M.	0,048344	0,000090	11				
Li, Xiao	0,000126	0,000000	9	Li, X.P.	0,061371	0,000026	15				
Li, Xiao-gang	0,000090	0,000000	5	Li, X.Q.	0,061130	0,000422	15				
Li, Xiaohong	0,000110	0,000000	7	Li, X.X.	0,000116	0,000000	05				
Li, Xiaokun	0,000107	0,000000	4	Li, X.Y.	0,067275	0,002028	31				
Li, Xiao-ping	0,007970	0,000000	13	Li, X.Z.	0,056115	0,000000	07				
Li, Xiaoqian	0,000185	0,000000	6		.						
Li, Xiaoyun	0,000121	0,000000	4								
Li, Xin	0,000059	0,000000	3								
Li, Xing	0,000069	0,000000	4								
Li, Xinhu	0,000096	0,000000	6								
Li, Xin-Hu	0,000055	0,000000	3								
Li, Xiuxi	0,000095	0,000000	5								
Li, Xiu-ying	0,007970	0,000000	13								
Li, X.J.	0,000055	0,000000	3								
Li, X.M.	0,000110	0,000000	7								
Li, Xuanqi	0,007900	0,000011	9								
Li, Xue	0,000069	0,000000	4								
Li, Xueping	0,000041	0,000000	2								
Li, Xueyan	0,000069	0,000000	4								
Li, X.Y.	0,000180	0,000000	6								

Fonte: Elaborada pelos autores (2014).

Notas: *Bet.*: *Betweenness*; *Close.*: *Closeness*; Rede I: Nome Completo; Rede II: Nome abreviado; Rede III: Sobrenome acompanhado da primeira inicial.

http://dx.doi.org/10.1590/0103-37862015000300001

vértices. Assim, se as distâncias geodésicas diminuem, a medida de proximidade (*closeness*) aumenta.

Quanto ao *betweeness,* como essa medida de centralidade atribui importância a um ator em função da passagem de fluxo por ele, quando os autores são agrupados, aumenta a probabilidade de que um fluxo maior de informações passe pelo novo vértice. Dessa forma, já era esperado que a medida de *betweeness* do autor "Folke, C" aumentasse na Rede III quando comparada com as Redes I e II, e que os autores mais centrais da Rede III possuíssem valores de intermediação mais altos que os das demais redes.

De acordo com a Tabela 1, o autor "Li, X", presente apenas nos resultados da Rede III, foi o que apresentou os maiores valores de proximidade (*closeness*) e intermediação (*betweeness*), figurando também entre os 10 mais centrais no *ranking* da medida de grau. No entanto, conforme mostrado na Figura 1, o autor que figura como "Li, X" na Rede III representa a contração de 30 nomes completos considerados na Rede I e de 15 nomes abreviados considerados na Rede II. Para verificar o posicionamento de "Li, X" nas demais redes, foi feito o cálculo das medidas de grau, *closeness* e *betweeness* sem a contração dos nomes, conforme listado na Tabela 2.

Os resultados da Tabela 2 mostram como as três medidas de centralidade de "Li, X" aumentaram da Rede I para a Rede II e da Rede II para a Rede III, em função do agrupamento de nomes de diferentes autores em um único vértice da rede.

Os resultados da aplicação da Correlação de Spearman entre as Redes podem ser encontrados na Tabela 3. Os valores de Correlação encontrados mostram que, mesmo havendo distorções nas posições relativas de correlação entre os autores, se analisada a distribuição dos valores de centralidade como um todo, a mudança de nomes afeta a centralidade de maneira semelhante à verificada na análise de redes com erros amostrais (Borgatti, 2006; Wang, 2012; Smith, 2013).

Discussão

Os resultados da comparação entre as Redes I, II e III mostram que as medidas de centralidade de grau, *closeness* e *betweeness* são bastante alteradas nos vértices principais em função da adoção de formas distintas de se considerar o nome dos autores na construção de uma rede de coautoria. Foi verificada também uma correlação de Spearman positiva entre as distribuições de centralidade entre os vértices. As diferenças observadas nos valores de cada medida acabaram por modificar também o ordenamento da importância atribuída a cada autor dentro das redes analisadas, de modo que cada rede apresentou um *ranking* diferente.

De acordo com o estudo desenvolvido, a rede construída a partir de uma forma mais simplificada de identificação dos autores, apenas sobrenome acompanhado da primeira inicial (Rede III), foi a que apresentou maiores discrepâncias quando comparada às demais. Na Rede III, o *ranking* das três medidas de centralidade consideradas foi composto, quase que exclusivamente, de autores com nomes de origem oriental. Esse fenômeno não aconteceu nas duas outras redes. Outro aspecto que chama atenção é que apenas um autor foi apontado como um dos mais centrais nas três redes. Esses resultados, portanto, sugerem uma significativa ocorrência de variações provocadas por ambiguidades de nomes, gerando os problemas apontados por Wang *et al*. (2012), citados na Introdução do presente artigo.

Os resultados de correlação são semelhantes aos verificados em outros estudos que tratam especificamente dos erros em redes (ausência de vértices e arestas), como os de Borgatti (2006), de Smith e Moody (2013) e

Tabela 3. Correlação de *Spearman* entre os vetores de centralidade nas três redes.

Redes	Grau	*Closeness*	*Betweenness*
Rede I - Rede II	0,871	0,759	0,945
Rede I - Rede III	0,717	0,580	0,917
Rede II - Rede III	0,852	0,768	0,943

Fonte: Elaborada pelos autores (2014).

Notas: Rede I: Nome Completo; Rede II: Nome abreviado; Rede III: Sobrenome acompanhado da primeira inicial.

de Costenbader e Valente (2003). É importante ressaltar que, em uma análise de coautoria, o estudo das distribuições de centralidade pode ser interessante para determinar a concentração das interações entre os membros. No entanto, as grandes diferenças entre os *rankings* levam a crer que a simples indicação dos autores com maior centralidade pode ficar bastante distorcida.

Os achados deste trabalho também reforçam a preocupação com a ambiguidade de nomes provocada pelo aumento da produção científica de países orientais, como a China, cujos autores apresentam pequena diversidade de sobrenomes, dificultando bastante a identificação correta do pesquisador (Tang & Walsh, 2010). Conforme exemplificado na Figura 1, um dos nomes relacionados no *ranking* das medidas de centralidade da Rede III resultou da contração do nome completo de 30 diferentes autores e do nome abreviado de 15, o que evidencia um falso agrupamento de vértices, provocando graves distorções na estrutura e nas métricas dessa rede.

As Redes I e II, por sua vez, também apresentaram variações, corroborando a existência de problemas mencionados na literatura sobre o assunto (Smalheiser & Torvik, 2009), tais como: existência de homônimos; mudanças de nomes; e registro de nomes completos e abreviados de forma diferente nas várias publicações de um mesmo autor. Esse é um aspecto a ser levado em consideração pelos pesquisadores, que devem ser orientados a manter a mesma forma do registro do nome em todas as publicações. É preciso ter ciência de que variações na grafia do nome aumentam a incidência de erros na recuperação e uso de dados bibliométricos, o que pode interferir nas estatísticas de produção científica, tanto do pesquisador quanto da instituição à qual ele está vinculado.

Considerando que a produção científica no mundo e o volume de informações disponibilizadas nas diversas bases de dados vêm aumentando consideravelmente, o tratamento adequado de normalização dos nomes dos autores se torna cada vez mais premente na construção de redes de coautoria. Os resultados encontrados mostram quanto as métricas dessas redes são sensíveis a problemas provocados pela ambiguidade dos nomes. Apesar da importância do tema no contexto dos estudos de ARS, existem poucos trabalhos que buscam quantificar o impacto provocado por esse tipo de erro nas medidas comumente utilizadas para analisar as redes sociais.

Apesar de diversos trabalhos proporem soluções para mitigar esse problema através de diversos métodos computacionais, trata-se de uma questão ainda em aberto e que vem se tornando cada vez mais complexa. Na literatura sobre Bancos de Dados e Sistemas de Informação, são encontrados diversos métodos que buscam reduzir problemas decorrentes da ambiguidade dos nomes (Ferreira *et al*., 2012; Peng *et al*., 2012; Sun *et al*., 2013), muitos dos quais baseados em mineração de texto e no processamento de linguagem natural (Amancio *et al*., 2012). No entanto, conforme apontado por Tang e Walsh (2010), trata-se de um problema ainda não resolvido e que requer maior aprofundamento. Uma alternativa que vem sendo levada em consideração, apresentada como serviço pelos administradores de bases de dados bibliográficos, é o registro único de cada autor, como, por exemplo, o *Research ID* e o *Open Researcher and Contributor ID* (ORCID), que funcionariam de forma similar ao *Digital Object Identifier* (DOI) de um artigo publicado.

No entanto, é importante salientar que, enquanto não se tem uma solução para esse problema, os estudiosos de ARS devem adotar procedimentos meticulosos na etapa de normalização de nomes. Do contrário, os resultados encontrados não corresponderão à realidade, podendo levar a falsas considerações e conclusões. Se isso acontecer, ficará comprometido o objetivo de se utilizar a ARS com o propósito de identificar e analisar as interações e relações entre as pessoas de modo a compreender a estrutura relacional de um determinado grupo social (Marteleto & Tomaél, 2005; Cruz, 2010).

Conclusão

Este estudo teve como objetivo mostrar as distorções que acontecem em métricas de centralidade, que visam identificar a importância de um ator dentro de uma rede social, em razão da ambiguidade de nomes. Para esse propósito foram construídas redes de coautoria, fazendo uso de três diferentes formas de registro do

http://dx.doi.org/10.1590/0103-37862015000300001

nome dos autores. Os resultados apontaram grandes variações no posicionamento dos autores dentro das redes analisadas.

Apesar da relevância do assunto, existem poucos trabalhos que buscam quantificar quanto os erros provocados pela ambiguidade dos nomes interferem nas métricas das redes. Apesar das soluções computacionais que têm sido propostas para mitigar o problema, trata-se de uma questão ainda não equacionada e que demanda aprofundamento.

Como erros dessa natureza podem provocar distorções significativas nos resultados, é importante que os pesquisadores estejam cientes de que é preciso adotar procedimentos cuidadosos na etapa de normalização de nomes, bem como descrever de forma clara e detalhada os procedimentos adotados, a fim de que os achados de suas pesquisas possam ser validados pelos pares.

Portanto, diante da proliferação dos estudos de ARS e do aumento da dimensão e da complexidade das redes, o presente estudo chama atenção para a necessidade de maior discussão do problema no ambiente da academia, em busca de alternativas de solução.

Agradecimentos

À Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (Processo nº E-26/111.753/2011) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

Referências

Abbasi, A.; Altmann, J.; Hwang, J. Evaluating scholars based on their academic collaboration activities: two indices, the RC-index and the CC-index, for quantifying collaboration activities of researchers and scientific communities. *Scientometrics*, v.83, n.1, p.1-13, 2010.

Acedo, F.J. *et al*. Co-Authorship in management and organizational studies: An empirical and network analysis. *Journal of Management Studies*, v.43, n.5, p.957-983, 2006.

Amancio, D.R.; Oliveira Jr., O.N.; Costa, L.F. On the use of topological features and hierarchical characterization for disambiguating names in collaborative networks. *Europhysics Letters*, v.99, n.4, p.1-6, 2012.

Barabasi, A.L. *et al*. Evolution of the social network of scientific collaborations. *Physica A*, v.311, n.3-4, p.590-614, 2002.

Barbastefano, R.G. *et al*. Impactos dos nomes nas propriedades de redes sociais: um estudo em rede de coautoria sobre sustentabilidade. *Perspectivas em Ciência da Informação*, v.18, n.3, p.78-95, 2013.

Borgatti, S.P.; Carley, K.M.; Krackhardt, D. On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, v.28, n.2, p.124-136, 2006.

Çavusoglu, A.; Türker, I. Scientific collaboration network of Turkey. *Chaos, Solitons & Fractals*, v.57, p.9-18, 2013.

Costenbader, E.; Valente, T.W. The stability of centrality measures when networks are sampled. *Social Networks*, v.25, n.4, p.283-307, 2003.

Cruz, R.C. Redes sociais virtuais: premissas teóricas ao estudo em ciência da informação. *TransInformação*, v.22, n.3, p.255-272, 2010.

De Souza, C.G.; Barbastefano, R.G.; De Lima, L.S. Chemistry colaboration networks in Brazil: A coauthorship study in quimica nova articles. *Química Nova*, v.15, n.4, p.671-676, 2012.

Ferreira, A.A. *et al*. A tool for generating synthetic authorship records for evaluating author name disambiguation methods. *Information Sciences*, v.206, p.42-62, 2012.

Freeman, L. Centrality in social networks. *Social Networks*, v.1, p.215-239, 1979.

Glanzel, W. Coauthorship patterns and trends in the sciences (1980-1998): A bibliometric study with implications for database indexing and search strategies. *Library Trends*, v.50, n.3, p.461-473, 2002.

Jiang, Y. Locating active actors in the scientific collaboration communities based on interaction topology analyses. *Scientometrics*, v.74, n.3, p.471-482, 2008.

Kang, I-S. *et al*. On co-authorship for author disambiguation. *Information Processing and Management*, v.45, p.84-97, 2009.

Kossinets, G. Effects of missing data in social networks. *Social Networks,* v.28, p.247-268, 2006.

Li-Chun, Y. *et al*. Connection and stratification in research collaboration: An analysis of the COLLNET network. *Information Processing and Management*, v.42, p.1599-1613, 2006.

Marteleto, R.M.; Tomaél, M.I. A metodologia de análise de redes sociais (ARS). In: Valentim, M.L.P. (Org.). *Métodos qualitativos de pesquisa em ciência da informação*. São Paulo: Polis, 2005. p.81-100.

Newman, M.E.J. Scientific collaboration networks I: Network construction and fundamental results. *Physical Review E*, v.64, n.1, p.1-8, 2001.

Nooy, W.; Mrvar, A.; Batagelj, V. *Exploratory network analysis with pajek*. Cambridge: Cambridge University Press, 2005.

Palla, G.; Barabasi, A.L.; Vicsek, T. Quantifying social group evolution. *Nature*, v.446, n.7136, p.664-667, 2007.

Peng, H-T. *et al.* Disambiguating authors in citations on the web and authorship correlations. *Expert Systems with Applications*, v.39, n.12, p.10521-10532, 2012.

Smalheiser, N.R.; Torvik, V.I. Author name disambiguation. *Annual Review of Information Science and Technology*, v.43, n.3, p.287-313, 2009.

Smith, J.A.; Moody, J. Structural effects of network sampling coverage I: Nodes missing at random. *Social Networks*, v.35, n.4, p.652-668, 2013.

Sun, X. *et al*. Ambiguous author query detection using crowdsourced digital library annotations. *Information Processing and Management*, v.49, n.2, p.454-464, 2013.

Tang, L.; Walsh, J.P. Bibliometric fingerprints: Name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, v.84, n.3, p.763-784, 2010.

Tomaél, M.I.; Marteleto, R.M. Redes sociais de dois modos: aspectos conceituais. *Transinformação*, v.25, n.3, p.245-253, 2013.

Wang, D.J. *et al*. Measurement error in network data: A re-classification. *Social Networks*, v.34, p.396-409, 2012.

Wasserman, S.; Faust, K. *Social network analysis*: *Methods and applications*. New York: Cambridge University Press, 1994.

Yan, E.; Guns, R. Predicting and recommending collaborations: an author-, institution-, and country-level analysis. *Journal of Informetrics*, v.8, n.2, p.295-309, 2014.

http://dx.doi.org/10.1590/0103-37862015000300001