Impactos dos nomes nas propriedades de redes sociais: um estudo em rede de coautoria sobre sustentabilidade

Rafael Garcia Barbastefano

Doutor em Engenharia de Produção pela UFRJ. Professor Associado do CEFET/RJ

Cristina Souza

Doutora em Engenharia de Produção pela UFRJ Professora Associada do CEFET/RJ

Juliana de Sousa Costa

Mestranda do Programa de Pós-Graduação em Tecnologia do CEFET/RJ

Patrícia Mattos Teixeira

Mestranda do Programa de Pós-Graduação em Tecnologia do CEFET/RJ

A identificação correta dos autores é fator crítico em estudos de Análise de Redes Sociais (ARS) que envolvem redes de coautoria. O objetivo do trabalho é apresentar as diferenças de propriedades entre três redes de coautoria geradas com formas distintas de considerar o nome dos autores. Foi adotado o método bibliométrico. As redes foram construídas com base em 28.916 artigos sobre sustentabilidade, indexados no ISI/Web of Science, fazendo uso do software Pajek. Na análise, foram comparadas as seguintes propriedades: densidade, grau médio, componente gigante, distribuição dos graus, distância média, diâmetro da rede e coeficiente de clusterização de Watts-Strogatz. Os resultados indicaram grandes diferenças entre as redes, sugerindo que estudos de coautoria, fazendo uso de ARS, podem ter resultados comprometidos, caso não haja o tratamento adequado dos nomes dos autores. Diante do crescimento das aplicações de ARS, faz-se necessário o desenvolvimento de estudos e ferramentas voltados para mitigar a ocorrência desses erros.

Palavras chave: *Análise de redes sociais; Coautoria; Desambiguação; Sustentabilidade; Web of Science.*

Names and its impacts on social networks properties: a study in a co-authorship network on *sustainability*

The authors' correct identification is a critical factor in studies of Social Network Analysis (SNA) involving co-authorship networks. The objective of this paper is to present the differences in properties between three co-authorship networks, generated from 28,916 articles on sustainability indexed in ISI/Web of Science, with different methods of identifying the authors. Density, diameter, average degree, Watts-Strogatz clusterization coefficient, giant component and the degree distributions were analyzed. The results indicated significant differences among networks, suggesting that studies based on co-authorship networks involving SNA may have impaired results if a proper treatment of authors' names is not performed.

Keywords*: Social Network Analysis; Co-authorship; Disambiguation; Sustainability; Web of Science.*

Recebido em 03.04.2013 Aceito em 05.06.2013

1 Introdução

A Análise de Redes Sociais (ARS) é tema que tem despertado grande interesse, observando-se o crescente número de artigos com aplicações nas mais diversas áreas e diferentes propósitos. Dentre os estudos de redes sociais, uma abordagem recorrente na literatura é a análise das redes de coautoria que, segundo Abbasi, Altmann, Huang (2010), é uma das maneiras mais visíveis e acessíveis para identificar as relações de colaboração científica no ambiente da academia. Vanz e Stumpf (2010) também dizem que, apesar de suas limitações, a coautoria vem sendo empregada com sucesso na área de bibliometria, para investigar as relações colaborativas entre pessoas, instituições e países.

Uma rede de coautoria pode ser representada através de um grafo em que os vértices representam os autores e as arestas representam as publicações em parceria. Vários pesquisadores têm desenvolvido trabalhos, buscando melhor compreender a evolução e estrutura dessas redes, fazendo uso de ARS (NEWMAN, 2001; GOYAL; VAN DER LEIJ;

MORAGA-GONZALEZ, 2006; GOLDENBERG *et al.*, 2010; SOUZA; BARBASTEFANO, 2011).

No entanto, um grave problema apontado na literatura, em relação à estruturação de redes sociais baseadas em coautoria, é a possibilidade de erros decorrentes de ambiguidades no nome dos autores (MOODY, 2004; NEWMAN, 2004; SMALHEISER; TORVIK, 2009; MILOJEVIC, 2010). A existência de homônimos, grafias diferentes, mudanças de nomes, nomes incompletos e abreviações são ocorrências que podem prejudicar os resultados encontrados em estudos dessa natureza. Como exemplo, pode ser citado o caso de um famoso cientista, na área de ciência da computação, chamado Jeffrey D. Ullman, da Universidade de Stanford, cujo nome aparece com 10 variações incorretas na listagem de autores do Portal da ACM (ON, 2008).

O tratamento dos problemas de nomes em bases de dados possibilitou o aparecimento de diversos trabalhos, propondo metodologias e algoritmos que visam separar nomes grafados da mesma maneira ou unificar nomes redigidos de formas diferentes (ON, 2008; MINH VU; TAKASU; ADACHI, 2008; KANG *et al.*, 2009; KANG *et al.*, 2011; FERREIRA *et al.*, 2012a; PENG *et al.*, 2012; SUN *et al.*, 2013). Trata-se, entretanto, de um problema ainda não equacionado, que pode gerar muitas distorções. Afinal, qual o impacto da utilização de nomes incorretos na elaboração de uma rede de coautoria?

Buscando evidenciar as consequências desse problema, o objetivo do trabalho é apresentar as diferenças entre três redes de coautoria, construídas a partir de métodos distintos de identificação dos autores, sem o devido tratamento de redução de ambiguidades. Essas redes foram elaboradas, considerando-se: (i) nome completo; (ii) nome abreviado; e (iii) sobrenome acompanhado da primeira inicial do nome dos autores. Com base nas relações de coautoria de 28.916 artigos sobre sustentabilidade recuperados na base ISI/*Web of Science*, os resultados dessas três redes foram comparados em relação às seguintes métricas e propriedades: densidade, grau médio, componente gigante, distribuição dos graus, distância média, diâmetro da rede e coeficiente de clusterização de Watts-Strogatz.

Como a identificação correta do nome de autores é uma questão ainda não resolvida, esse estudo reforça a importância de ampliar a discussão sobre o assunto, visando encontrar soluções capazes de mitigar as consequências desse problema. O estudo também contribui para alertar os pesquisadores que atuam com ARS sobre o cuidado que se deve ter com o tratamento do nome dos autores, bem como sobre a necessidade de descrever claramente o método adotado nas diversas aplicações realizadas.

O trabalho encontra-se organizado em seções. As seções 2 e 3 trazem uma breve revisão da literatura, abordando o problema da normalização dos nomes em estudos de coautoria e a definição das propriedades de redes utilizadas nesse trabalho. A seção 4 descreve a

metodologia utilizada. Os resultados são apresentados e discutidos na seção 5. Segue a seção 6 com as conclusões.

2 Coautoria e normalização dos nomes dos autores

A prática de escrever artigos em coautoria se tornou mais comum nas últimas décadas (ACEDO *et al.*, 2006). Trata-se de um fenômeno social que vem acontecendo nas diversas áreas do conhecimento e cujo crescimento ocorre tanto em incidência (fração de artigos que apresentam coautoria) quanto em extensão (quantidade de coautores nas publicações) (LABAND; TOLLISON, 2000).

Como razões para o aumento da realização de trabalhos em coautoria, podem ser citadas: as novas Tecnologias da Informação e Comunicação (TIC), que favorecem o trabalho colaborativo à distância (LABAND; TOLLISON, 2000); políticas governamentais e agências de fomento, que estimulam a cooperação interinstitucional e internacional (KRETSCHMER, 2004; LEE; BOZEMAN, 2005); os altos custos de Pesquisa e Desenvolvimento (P&D) que fazem com que pesquisadores compartilhem recursos e infraestrutura; a necessidade de especialização, principalmente nas áreas em que a instrumentalização é complexa, fazendo com que a colaboração ocorra em função da necessidade de divisão do trabalho; e a interdisciplinaridade da ciência, que demanda pesquisadores advindos de diferentes áreas de conhecimento (LEE; BOZEMAN, 2005; MATHEUS; VANZ; MOURA, 2007).

De acordo com Maia e Caregnato (2008), à medida que as publicações em coautoria foram aumentando, o interesse em entender essas relações de colaboração também cresceu. Segundo as autoras, houve, ainda, a intensificação dos estudos de ARS, que proporcionam uma compreensão mais ampla das interações entre as partes, diferentemente de outras abordagens com foco nas características dos indivíduos. Como ressaltado por Matheus e Silva (2006, p. 1), a "diferença fundamental da ARS para outros estudos é que a ênfase não é nos atributos (características) dos atores, mas nas ligações entre os elos; ou seja, a unidade de observação é composta pelo conjunto de atores e seus laços".

Nos estudos de ARS, como a análise recai sobre a rede, a definição correta dos atores (vértices) e dos laços (arestas) que os unem se torna fator crítico para a obtenção de resultados corretos e confiáveis. Essa questão vem sendo bastante discutida na literatura. Marsden (1990), por exemplo, diz que nas redes sociométricas existe preocupação com a perda de dados decorrentes dos métodos de coleta utilizados, normalmente *surveys* e questionários, bem como de limitações cognitivas dos informantes. Segundo o autor, a acurácia com que as pessoas podem prover dados sobre seus laços relacionais não é uma questão trivial. Por outro lado, em redes complexas, que abrangem grande quantidade de vértices e relações e que são construídas com o apoio de recursos computacionais, o problema passa a ser a abundância de dados (DE CHOUDHURY *et al.*, 2010).

As redes de coautoria se enquadram nesse segundo caso, podendo ser caracterizadas como redes complexas, envolvendo milhares de vértices e cuja obtenção e tratamento de dados vêm acontecendo de forma automatizada. Essas redes estão sujeitas aos seis tipos de erros classificados por Wang *et al.* (2012a): (i) falsos vértices negativos – ausência de vértices que deveriam compor a rede; (ii) falsos vértices positivos – vértices erroneamente inseridos na rede; (iii) falsas arestas negativas – ausências de ligações entre vértices que deveriam estar presentes; (iv) falsas arestas positivas – ligações erroneamente representadas na rede; (v) vértices falsamente agregados – situação em que dois ou mais vértices são erroneamente considerados como um único vértice; e (vi) vértices falsamente desagregados – situação em que um vértice é erroneamente considerado como dois ou mais vértices separados.

A identificação correta dos autores, entretanto, não é problema de fácil solução. Conforme apontado em diversos trabalhos (KNAG *et al.*, 2009; WANG *et al.*, 2012a), um mesmo autor pode ter múltiplos nomes decorrentes de abreviações, omissões, mudanças de nome, pseudônimos e erros ortográficos, ao passo que diferentes autores podem se apresentar com o mesmo nome (homônimos). Casos como esses podem gerar erros de vértices falsamente agregados e desagregados.

A questão se torna mais crítica se considerarmos que alguns países, que vem se destacando no cenário internacional de produção científica, possuem sobrenomes bastante comuns, como é o caso da China (TANG; WALSH, 2010). Ressalta-se, ainda, que as bases de dados bibliográficas, muitas vezes, fazem registros dos metadados de forma incompleta (SMALHEISER; TORVIK, 2009), o que pode provocar distorções.

O problema de ambiguidade de nomes também se faz presente em outras atividades, como em análise de sistemas de informação, de modo que diversos pesquisadores têm desenvolvido métodos para tentar mitigá-lo. Peng *et al.* (2012) e Ferreira *et al.* (2012a) relacionam vários trabalhos encontrados na literatura que apresentam esse propósito. De acordo com Ferreira; Machado; Gonçalves (2012b), os métodos mais eficazes têm sido aqueles que tentam associar um conjunto de registros aos autores, de modo que esses não sejam considerados apenas por seus nomes, através de técnicas de aprendizado de máquina supervisionado. Amancio; Oliveira; Costa (2013), por sua vez, dizem que esses métodos têm sido baseados principalmente em mineração de texto e no processamento de linguagem natural, baseando-se na ideia de que medidas de similaridade textuais são capazes de agrupar manuscritos de autoria do mesmo cientista.

Wang *et al.* (2012b) também citam estudos voltados para o tratamento da ambiguidade de nomes. Segundo os autores, esse problema vem sendo tratado geralmente de duas formas. A primeira como um problema de clusterização, uma segunda como um problema de classificação binária, na qual podem ser aplicadas técnicas baesianas.

Smalheiser e Torvik (2009) tratam soluções do ponto de vista da aplicação de técnicas de *data mining*.

Smalheiser e Torvik (2009) também indicam a possibilidade de solução através da identificação única de autores. Seria um identificador para autores, equivalente ao doi. (*Digital Object Identifier* - http://doi.dx.org). Tal abordagem poderia resolver o problema de ambiguidade no longo prazo, caso uma solução universal fosse adotada.

Apesar dos esforços empreendidos para minimizar erros decorrentes da identificação incorreta dos autores, Tang e Walsh (2010) dizem que essa questão não tem sido suficientemente explorada na área da bibliometria e, muito menos, em análises de redes de coautoria. Segundo eles, alguns estudos evitam fazer análises no nível micro enquanto outros não apresentam o método, especificando claramente como esse problema foi tratado ou simplesmente mostram os resultados das análises mantendo a identificação dos autores como uma "caixa preta". Diante desse contexto, o presente trabalho procura mostrar como a falta de tratamento adequado do nome dos autores pode afetar a estrutura de uma rede de coautoria, impactando em diversas métricas de análise de redes sociais.

3 Métricas de Análise de Redes Sociais

Uma rede social é um Grafo G(V,E), onde V é o conjunto de vértices, que representam atores sociais e E o conjunto de arestas que representam relações sociais (WASSERMAN; FAUST, 1994; SCOTT, 2000). No caso de uma rede de coautoria, os vértices representam os autores, que são ligados por arestas, caso tenham produzido um artigo em comum

As características morfológicas de uma rede podem ser identificadas através de medidas de ARS. Neste estudo, serão abordadas as propriedades descritas a seguir: densidade, grau médio, distância média, diâmetro da rede e coeficiente de clusterização de Watts-Strogatz, componente gigante, lei de potência.

> *a)densidade* – medida que avalia a coesão das redes. A densidade é calculada pela razão entre o número de arestas e o número máximo de arestas possível em uma rede. Quanto maior a densidade, maior será a coesão de uma rede;
>
> *b)grau médio* – medida de coesão da rede, assim como a densidade, e quanto maior for o grau médio, maior será a coesão da rede. O grau mede o número de ligações de um vértice, o grau médio é a média de ligações de todos os vértices;
>
> *c)distância média* – medida que está relacionada ao número de arestas mínimo entre os caminhos que ligam dois vértices (distância dos vértices). A média das distâncias dos vértices tomados dois a dois é a distância média do grafo;

d)diâmetro – medida que calcula a maior distância possível entre dois vértices em uma componente conexa de um grafo;

e)em redes de coautoria, pode-se verificar a hipótese do mundo pequeno, a qual indica que a distância média entre os membros de determinadas redes é pequena. De acordo com essa hipótese, uma distância média pequena entre os vértices, por volta de seis, significa que a comunidade representada pelo grafo se ajusta a tal hipótese (MILGRAM, 1967; NEWMAN, 2010). O interesse principal desse estudo não era somente analisar o tamanho dessas ligações, mas, também, as características dos atores intermediários que fazem parte dessa rede (WASSERMAN; FAUST, 1994);

f)coeficiente de clusterização de Watts-Strogatz – medida do quanto um grafo se aproxima localmente de um grafo completo ou clique. O coeficiente de clusterização mede a razão média entre o número de pares de vértices conectados por arestas e o número de pares de vizinhos. (WATTS; STROGATZ, 1998);

lei de potências – propriedade relacionada à distribuição de graus. Diz-se que uma rede segue uma distribuição de lei de potência quando a distribuição dos graus (o número de ligações que um vértice possui com outros vértices) segue uma expressão do tipo:

$f(x) = k\, x^{-\alpha}$.

g)Na função f(x), que fornece o número esperado de vértices com grau x, o coeficiente k representa o número de vértices com grau 1; □ é uma constante; e x representa o grau dos vértices. Redes que seguem essa lei apresentam um padrão em que poucos autores possuem muitas conexões enquanto a maioria deles está ligada a apenas alguns autores. Nessas redes, a distribuição dos graus não se acumula em torno de uma média e, por isso, também são chamadas redes livres de escala; e

h)componente gigante – Propriedade que consiste no maior subgrafo conexo de uma rede. Em algumas redes de coautoria, a componente gigante pode abranger mais de 90% dos vértices da rede (NEWMAN, 2001).

4 Procedimentos metodológicos

Trata-se de uma pesquisa descritiva, com abordagem quantitativa (GIL, 1991). Foi adotado o método bibliométrico, sendo construídas três redes de coautoria, com o objetivo de comparar os resultados encontrados ao se utilizar os nomes dos autores de formas diferentes, sem fazer tratamento para remoção de ambiguidades. A primeira rede foi elaborada

a partir do nome completo dos autores. A segunda usou o nome abreviado, fornecido pelo ISI/*Web of Science*. Por fim, a terceira foi gerada com o sobrenome, acompanhado da primeira inicial do nome dos autores. As três redes foram construídas com base no mesmo conjunto de artigos.

Para a construção dessas redes, foi feito um levantamento dos artigos sobre o tema sustentabilidade publicados em periódicos indexados na ISI/*Web of Science*. Esse tema foi escolhido por ser atual, global e interdisciplinar, abrangendo pesquisadores de todo o mundo oriundos de diversas áreas do conhecimento. A escolha da *Web of Science* deveu-se aos seguintes fatores: trata-se de uma das mais importantes bases científicas do mundo. Ela é utilizada para geração de indicadores internacionais de produção científica, apresentando diversas estatísticas e grande parte dos estudos realizados, na área de bibliometria, faz uso dessa base para o levantamento de dados bibliográficos.

Como estratégia de busca, adotou-se o termo "*sustainability*" no campo título ou resumo, sendo excluídos artigos de conferências e editoriais. O levantamento foi realizado no mês de fevereiro de 2013 e foram considerados os documentos publicados até o final do ano de 2012. Como resultado, obteve-se um total de 28.916 artigos.

Cada registro possui o nome completo dos autores e o nome resumido, dentre outros campos. A partir do nome resumido, foi elaborada uma lista com o sobrenome mais a primeira inicial. Por exemplo: o autor "Tomich, Thomas P.", possui como nome resumido "Tomich, TP". Foi feito um tratamento, portanto, deste nome, para obtenção do sobrenome com a primeira inicial: "Tomich, P".

Com pares "Autor-Artigo", nos três formatos usados (nome completo, abreviado e sobrenome, primeira inicial), foram geradas três redes de autoria no formato do *software* Pajek , específico para análise de redes sociais (NOOY; MRVAR; BATAGELJ, 2005). O programa é gratuito e pode ser obtido em http://pajek.imfm.si. Em uma rede de autoria, cada vértice é um autor ou um artigo e artigos são ligados aos seus autores por arestas. É uma rede preparatória, obtida de modo mais direto do que a rede de coautoria.

Cada uma das três redes de coautoria foi gerada através do próprio programa Pajek, através da conversão das redes bipartidas de autoria em redes de coautoria. Os resultados das redes foram comparados em relação às seguintes propriedades e métricas, apresentadas na seção 3: densidade, grau médio, componente gigante, distribuição dos graus, distância média, diâmetro da rede e coeficiente de clusterização de Watts-Strogatz.

5 Resultados e discussão

Conforme apresentado na seção anterior, foram obtidas três redes de coautoria:

1) Rede com nomes completos dos autores fornecidos pelo ISI/*Web of Science* (Rede I);

2)Rede com nomes abreviados, fornecidos pelo ISI/*Web of Science* (Rede II); e

3)Rede com sobrenome e primeira inicial (Rede III).

Os resultados das análises serão discutidos ao longo desta seção e podem ser consultados na Tabela 1:

Tabela 1 – Resultados das análises

	Rede		
Propriedade	**Nome Completo (Rede I)**	**Nome Resumido (Rede II)**	**Sobrenome mais 1ª inicial (Rede III)**
Número de Vértices	72.808	64.051	55.908
Densidade	6,335E-05	8,054E-05	1,046E-04
Grau Médio	4,6125	5,1589	5,8485
Número de vértices na Componente Gigante	5.337	26.384	33.993
Percentual de vértices na Componente Gigante	7,3%	41,2%	60,8%
Coeficiente α da Lei de Potências	2,805	2,676	2,485
Coeficiente de Regressão da Lei de Potências - R^2	0,8848	0,9007	0,9240
Distância Média	9,568	9,014	6,742
Diâmetro	30	25	18
Clustering Coefficient (Watts-Strogatz)	0,9424	0,9025	0,8681

Fonte: Elaborado pelos autores.

A primeira informação que merece ser ressaltada se refere ao número de vértices das três redes. A Rede I possui o maior número de vértices e, conforme passamos para as Redes II e III, o número de vértices diminui. E essa diferença entre a quantidade de vértices ou tamanho da rede - a Rede I chega a ser 30% maior que a Rede III - pode ser atribuída à maneira como os nomes encontram-se em cada uma delas.

As abreviações, utilizadas nas Redes II e III, aumentam as chances de haver homônimos, isto é, autores que se juntem em um só vértice, devido à abreviação dos nomes. A Tabela 2 mostra um caso interessante. Um vértice na Rede III (Li, Y) pode representar 14 vértices diferentes na Rede II e 29 vértices na Rede I. É importante notar que mesmo o nome completo, no registro do ISI/*Web of Science,* pode aparecer resumido. Além disso, o uso do nome completo não implica necessariamente em uma rede sem erros, pois o mesmo autor pode ser representado por diversos nomes diferentes.

Tabela 2 – O caso do autor Li, Y

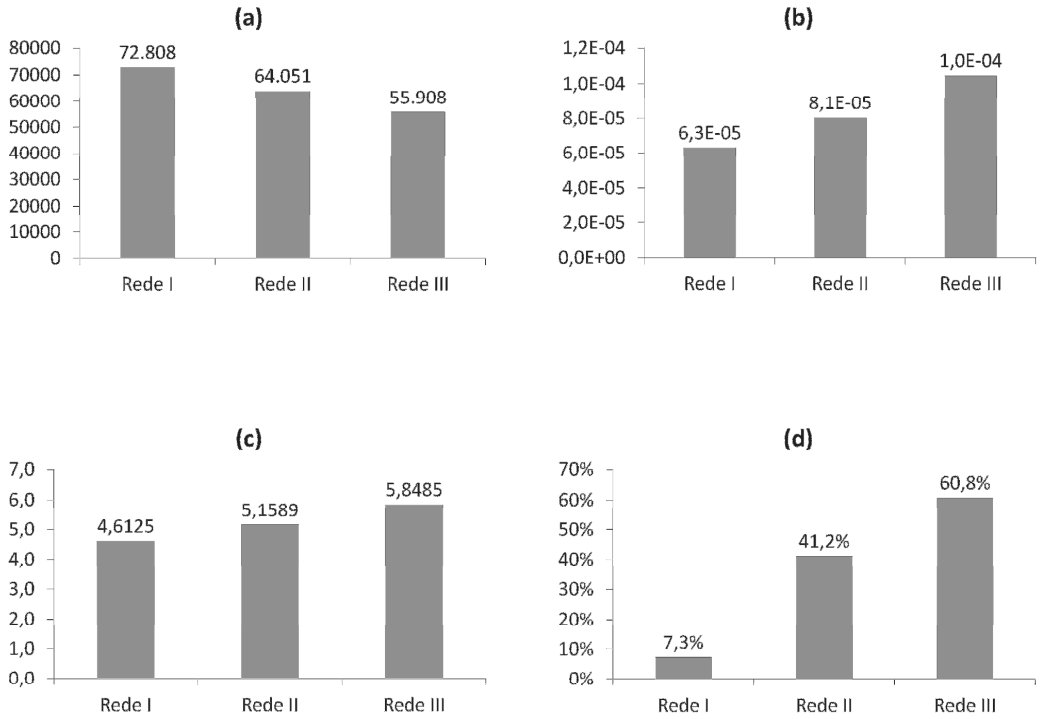
Nome Completo (29 nomes)			Nome Resumido (14 nomes)		Sobrenome mais 1ª inicial (1 nome)
Li, Y	Li, YC	Li, Yuan-Wei	Li, Y	Li, YT	Li, Y
Li, Y.	Li, YF	Li, Yuan-Yao	Li, YC	Li, YW	
Li, Y. P.	Li, Yiping	Li, Yue	Li, YD	Li, YX	
Li, Yadong	Li, Yong	Li, Yufei	Li, YF	Li, YY	
Li, Yan	Li, Yourun	Li, Yuhua	Li, YH		
Li, Yang	Li, YP	Li, Yuncong	Li, YM		
Li, Yangfan	Li, YR	Li, Yurui	Li, YP		
Li, Yan-Ping	Li, YS	Li, Yu-Yi	Li, YQ		
Li, Yaqiong	Li, Yuan	Li, YX	Li, YR		
Li, Yating	Li, YuanWei		Li, YS		

Fonte: Elaborado pelos autores

A junção de vértices, que passa da Rede I para a Rede II e Rede III, explica não só apenas a diferença significativa no tamanho das redes, mas, também, o aumento da densidade e do grau médio, assim como outros resultados desse estudo, como o percentual de vértices englobados pela componente gigante, registrado no gráfico da Figura 1d.

A densidade das redes aumenta em 65%, se considerarmos a rede da primeira inicial (Rede III) em comparação com a rede de nomes completos (Rede I). O mesmo resultado é esperado para a outra medida de coesão calculada, o grau médio. Dessa forma, tanto o grau médio quanto a densidade convergem para o entendimento que a Rede III é a mais coesa entre as três redes analisadas, seguida pela Rede II e, por fim, a Rede I, que é a menos coesa.

Figura 1 – (a) Número de vértices de cada rede; (b) Densidade de cada rede; (c) Grau médio de cada rede e (d) Percentual dos vértices englobado pela componente gigante de cada rede



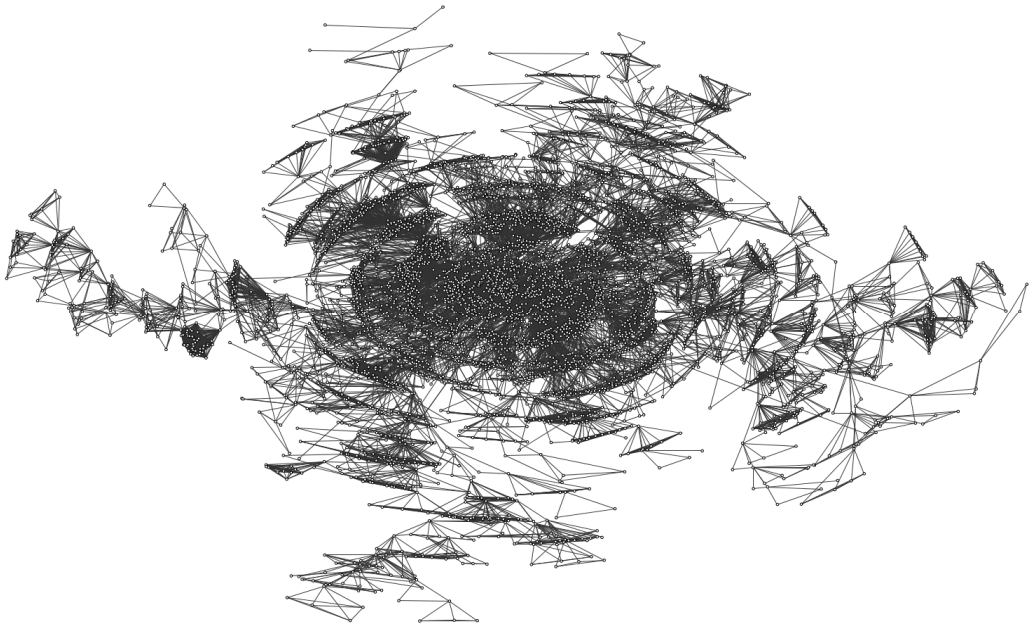
Fonte: Elaborado pelos autores.

Comparando, mais uma vez, a Rede III com a Rede I, verifica-se que o número de vértices que compõem a componente gigante aumenta mais de seis vezes, enquanto o percentual abrangido por ela aumenta mais de oito vezes. O resultado obtido é devido à abreviação do nome dos vértices, que antes pertenciam a diferentes grupos e passam a estar ligados e, dessa forma, formam um grupo ainda maior. A Figura 2 apresenta o exemplo da representação da componente gigante da Rede I. Conforme pode ser observado, essa componente apresenta alguns aspectos típicos de redes governadas por leis de potências, ou seja, possui *hubs* ao centro e vértices com um número pequeno de ligações nas extremidades.

Analisando a distribuição dos graus das redes, conforme pode ser visto na Figura 3, e os coeficientes de regressão obtidos pelo ajuste de uma lei de potências a cada uma das distribuições na Tabela 1, pode-se verificar aderência a uma lei de potências em cada uma das redes.

À medida que unimos os vértices, verificamos também uma redução dos coeficientes α nas distribuições dos graus, como se observa na Figura 3. Isto significa que as redes com nomes completos possuem maior concentração de vértices com muitas ligações, ou seja, proporcionalmente existem menos autores com maior grau. Já a Rede III, a distribuição de graus é menos concentrada, ou seja, ela possui, proporcionalmente, um número maior de *hubs* do que a Rede I, apresentando uma distribuição mais suave que as demais.

Figura 2 – Componente gigante da rede com nomes completos



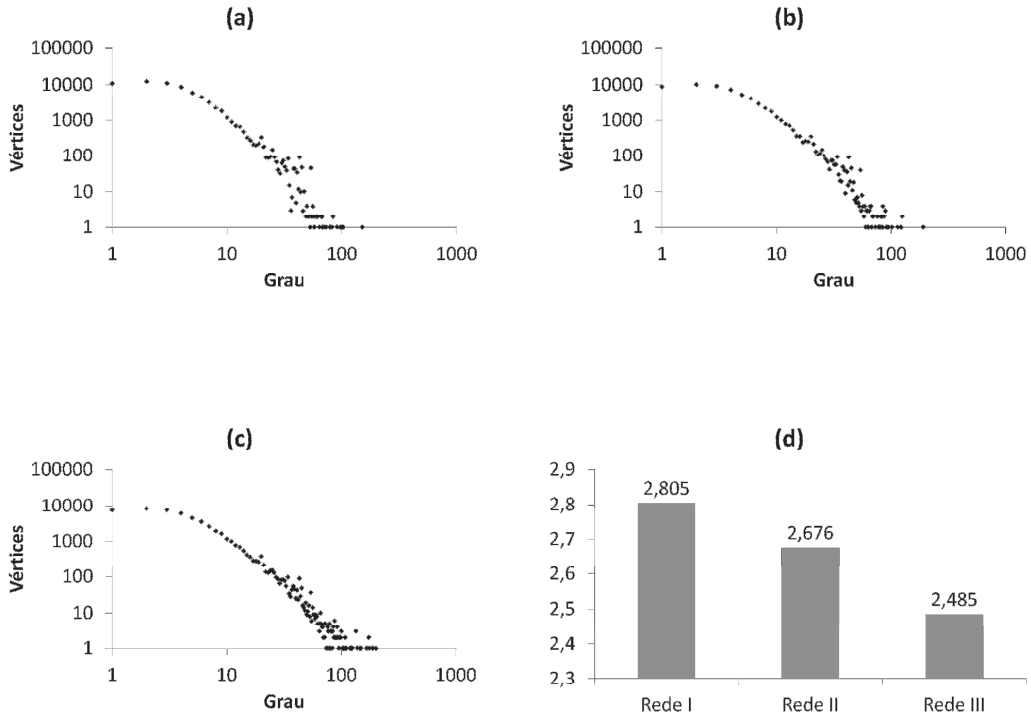
Fonte: Elaborado pelos autores

A distância média e o diâmetro também diminuíram com a junção dos vértices. Ao passar da Rede I, para as Redes II e III, caminhos alternativos são criados entre os vértices, reduzindo as distâncias entre eles. Além disso, pares de vértices desconectados na Rede I passaram a pertencer a uma mesma componente conexa nas Redes II e III.

É importante ressaltar que as distâncias pequenas indicam aderência à hipótese do mundo pequeno (MILGRAM, 1967; NEWMAN, 2010), citada na seção 4, sendo que, ao observar os resultados obtidos para cada uma das redes analisadas, é possível perceber que a Rede III se adequa ao valor de "seis graus de distância" entre os autores, verificado no próprio exemplo de Milgram.

Um resultado que chama a atenção é a diminuição dos coeficientes de clusterização de Watts-Strogatz (CC1), ao passar da Rede I para as Redes II e III. A princípio, o aumento de densidade e do grau médio favoreceriam um aumento e não uma redução do CC1. Assim, enquanto a densidade e grau médio aumentam, à medida que as redes passam a possuir menos vértices, os valores dos coeficientes de clusterização diminuem.

Figura 3 (a) Distribuição dos graus da Rede I; (b) Distribuição dos graus da Rede II; (c) Distribuição dos graus da Rede III; (d) Coeficiente α da lei de potências associada a cada rede



Fonte: Elaborado pelos autores.

Entretanto, apesar de aparentar ser um resultado contraditório, as características das redes de coautoria podem explicar a diminuição do CC1. Cada artigo em uma rede de coautoria é representado por um clique (grafo completo), já que todos os autores do mesmo artigo estão ligados. Em uma rede mais desconectada (Rede I, de nomes completos), a probabilidade de um vértice estar ligado a apenas cliques é maior do que em uma rede mais conectada.

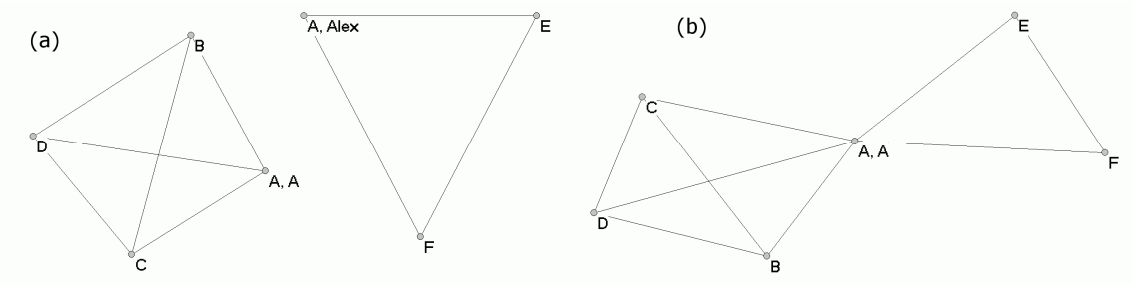
Na Figura 4, apresentamos um exemplo simples deste processo. A Figura 4a representa dois artigos, enquanto na Figura 4b, dois vértices separados são unidos. Os valores de densidade, grau médio e coeficiente de clusterização estão na Tabela 3

Figura 4 Exemplo de junção de vértices. Na rede (a), o coeficiente de clusterização é maior do que na rede (b). Entretanto, a rede (b) possui maior grau médio e densidade

Tabela 3 – Valores do exemplo

Rede	Vértices	Densidade	Grau Médio	CC1
Figura 4a	7	0,37	2,57	1,00
Figura 4b	6	0,50	3,00	0,90

Fonte: Elaborado pelos autores.



Fonte: Elaborado pelos autores.

Apesar da rede (b) ser mais densa e possuir grau médio maior do que a rede (a), quando juntamos o vértice "A, A" com o vértice "A, Alex", o vértice "A, A" na rede (b), localmente deixou de fazer parte de um clique, reduzido o CC1 médio. Tal situação se verifica nas redes estudadas, explicando a redução do coeficiente de clusterização, visto que, em uma rede desconexa, é mais fácil aparecerem ligações entre vértices que possuem um terceiro vértice em comum do que com um outro qualquer da rede. Já no caso em que a rede foi formada com nomes abreviados, como diversos autores estão representados pelo mesmo vértice, aumenta-se o número de ligações de cada um deles. E, assim, também a probabilidade de estarem ligados com diversos outros vértices, porém não apresentando estrutura de cliques, visto que esses outros vértices não necessariamente estão ligados uns aos outros.

6 Conclusão

A identificação correta dos autores é fator crítico nos estudos de ARS, que envolvem redes de coautoria. Trata-se de um problema ainda não equacionado, que pode gerar diversas distorções, levando à falta ou inclusão de vértices e arestas indevidas na rede, modificando a sua estrutura.

Este artigo teve por objetivo mostrar como os resultados das redes de coautoria variam de acordo com o método utilizado para considerar o nome dos autores sem o correto tratamento de remoção de ambiguidades.

Para o alcance desse objetivo, foram construídas três redes de coautoria, considerando-se o nome completo dos autores, o nome abreviado e o sobrenome acompanhado da primeira inicial. Essas redes foram construídas com base na autoria de 28.916 artigos sobre sustentabilidade, indexados na *Web of Science*. A comparação entre as redes foi feita a partir de diversas métricas de ARS.

Os resultados apontaram grandes diferenças entre as redes, sendo que as maiores variações aconteceram entre a rede de nome completo e a rede utilizando o sobrenome acompanhado da primeira inicial do nome dos autores. Enquanto a rede de nome completo apresentou 72.808 vértices, a rede do sobrenome com a primeira inicial teve apenas 55.908, ou seja, uma diferença de 16.900 vértices.

Essa falta de tratamento de ambiguidades entre os nomes dos autores fez com que o percentual de vértices abrangido pela componente gigante variasse de 7,3% a 60,8%; o diâmetro de 30 a 18; a distância média de 9,568 a 6,742; a taxa de decaimento nas leis de potência de 2,805 a 2,485. Através desses valores, pôde-se perceber o quanto a Rede I é mais concentrada que as demais, possuindo menor número de vértices com grande número de ligações. Além disso, outro resultado importante para avaliação e comparação dos resultados das três redes estudadas é o coeficiente de clusterização de Watts-Strogatz (CC1). O resultado deste mostrou que, apesar de, à primeira vista, poder deduzir-se que, conforme os valores de densidade e grau médio da rede aumentam, o CC1 também aumentaria, verifica-se que ocorre justamente o contrário. Este resultado, todavia, está de acordo com o que deve ser realmente esperado, visto que, quanto mais desconexa a rede, maior a probabilidade de existirem diferentes cliques, visto que muitos *clusters* serão justamente os artigos publicados.

Considerando-se as grandes variações entre as redes, a partir das métricas analisadas, pode-se dizer que os estudos de coautoria, fazendo uso de ARS, podem ter seus resultados comprometidos, caso não ocorra um tratamento adequado, visando à identificação correta dos autores. Porém, observa-se, também, que nem sempre tais tratamentos podem ser realizados de maneira simples. Isto porque, muitas vezes, a própria fonte de dados, como as bases em que são extraídos, têm os nomes dos autores indexados de forma não completa, como por abreviações ou até mesmo podem existir autores com nomes similares. Assim sendo, fica evidente a necessidade de que tais estudos especifiquem claramente o método utilizado, para reduzir ambiguidades e identificar corretamente os autores, de modo a garantir a transparência e credibilidade dos resultados encontrados.

Como possibilidade de estudos futuros, indica-se a verificação do impacto da utilização de ferramentas de desambiguação automática em métricas de redes para redução do problema indicado na Tabela 2. Por outro lado, ferramentas de *data warehouse,* como o Google Refine (http://code.google.com/p/google-refine/), podem ser usadas para tratamento do problema inverso, ou seja, a junção de vértices que representem o mesmo autor, grafado de duas maneiras diferentes, como, por exemplo, Tomich, T.P. e Tomich, TP.

Mesmo com tratamento automatizado de referências, ainda assim, a análise de redes bibliométricas de coautoria estará sujeita a erros. Soluções mais definitivas podem aparecer na aplicação de serviços de registro único de autores, como o ResearcherID da Thomson Reuters (http://www.researcherid.com) ou o ORCID (http://orcid.org).

Soluções de registro único de autores podem resolver o problema na teoria, entretanto, dependem da adesão voluntária dos autores e não resolvem o problema de ambiguidades de artigos passados, cujos autores fiquem sem indicação. Dessa maneira, é provável que tenhamos que conviver com o problema, descrito neste trabalho, ainda por vários anos.

7 Agradecimentos

Os autores gostariam de agradecer o apoio da Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) para a realização deste trabalho.

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, 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*, Oxford, v. 43, n. 5, p. 957-983, 2006.

AMANCIO, D. R.; OLIVEIRA JR.; O. N.; COSTA, L.; DA F. On the use of topological features and hierarchical characterization for disambiguating names in collaborative networks. *EPL*, v. 99, n. 4, 2012.

DE CHOUDHURY, M. *et al.* Inferring relevant social networks from interpersonal communication. *In:* INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 19., 2010, New York. *Proceedings...* New York: Association for Computing Machinery, 2010, p. 301–310.

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, 2012a.

FERREIRA, A. A.; MACHADO, T. M.; GONÇALVES, M. A. Improving Author Name Disambiguation with User Relevance Feedback. *Journal of Information and Data Management*, v. 3, n. 3, p. 332–347, 2012b.

GIL, A. *Como elaborar projetos de pesquisa*. São Paulo: Atlas, 1991.

GOLDENBERG, J. *et al.* The evolving social network of marketing scholars. *Marketing Science*, v. 29, p. 561-567, 2010.

GOYAL, S.; VAN DER LEIJ, M.; MORAGA-GONZALEZ, J. Economics: an emerging small world. *Journal of Political Economics*, v. 114, p. 403-412, 2006.

KANG, I-S. *et al.* Construction of a large-scale test set for author disambiguation. *Information Processing and Management*, v. 47, p. 452-465, 2011.

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

KRETSCHMER, H. Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the Web. *Scientometrics*, v. 60, n. 3, p. 409-420, 2004.

LABAND, D. N.; TOLLISON, R. D. Intellectual Collaboration. *Journal of Political Economy*, v. 108, n. 3, p. 632-662, 2000

LEE, S.; BOZEMAN, B. The Impact of research collaboration on scientific productivity. *Social Studies of Science*, v. 35, n. 5, p. 673-702, 2005.

MAIA, M. F. S.; CAREGNATO, S. E. Coautoria como indicador de redes de colaboração científica. *Perspectivas em Ciência da Informação*, v. 13, n. 2, p. 18-31, maio/ago. 2008.

MARSDEN, P. V. Network data and measurement. *Annual Review of Sociology,* v. 16, p. 435-463, 1990.

MATHEUS, R. F.; SILVA, A. B. O. S. Análise de redes sociais como método para a Ciência da Informação. *DataGramaZero - Revista de Ciência da Informação*, v. 7, n. 2, 2006.

MATHEUS, R. F.; VANZ, S. A.; MOURA, A. M. M. Coautoria e co-invenção: indicadores da colaboração em CT&I no Brasil. *In:* CONGRESO IBEROAMERICANO DE INDICADORES DE CIENCIA Y TECNOLOGÍA, 7., 2007, São Paulo. *Anais...* São Paulo: RICYT, maio de 2007.

MILOJEVIC, S. Modes of collaboration in modern science: beyond power laws and preferential attachment. *Journal of the American Society for Information Science and Technology*, v. 61, n. 7, p. 1410-1423, 2010.

MINH VU, Q.; TAKASU, A.; ADACHI, J. Improving the performance of personal name disambiguation using web directories. *Information Processing and Management*, v. 44, n. 4, p. 1546-1561, 2008.

MILGRAM, S. Small-world problem. *Psychology Today,* v. 1, p. 61-67, 1967.

MOODY, J. The structure of a social science collaboration network: disciplinary cohesion from 1963 to 1999. *American Sociological Review*, v. 69, n. 2, p. 213-238, 2004.

NEWMAN, M. E. J. The structure of scientific collaboration networks. *PNAS*, v. 98, p. 404-409, 2001.

NEWMAN, M.E.J. Who is the best connected scientist? A study of scientific coauthorship networks. *In:* BEN-NAIM, E.; FRAUENFELDER, H.; TOROCZKAI, Z. (Eds.). *Complex networks*. Berlin, Germany: Springer, 2004. p. 337-370.

NEWMAN, M. E. J. *Networks:* an introduction. New York: Oxford University Press, 2010.

NOOY, W; MRVAR, A.; BATAGELJ, V. *Exploratory Social Network Analysis with Pajek*. New York: Cambridge University Press, 2005.

ON, B-W. Social Network Analysis on Name Disambiguation and More. *In:* CONFERENCE ON CONVERGENCE AND HYBRID INFORMATION TECHNOLOGY, 3., 2008, Busan, CN. *Proceedings…* Busan, CN: IEEE, 2008. v. 2, p. 1081-1088.

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

SCOTT, J. *Social network analysis:* a handbook. 2. ed. London: Sage Publications, 2000.

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

SOUZA, C. G.; BARBASTEFANO, R. G. Knowledge diffusion and collaboration networks on life cycle assessment. *The International Journal of Life Cycle Assessment*, v. 16, n. 6, p. 561-568, 2011.

SUN, X. *et al.* Ambiguous author query detection using crowdsourced digital library annotations. *Information Processing and Management,* v. 49, 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.

VANZ, S. A. de S.; STUMPF, I. R. C. Colaboração científica: revisão teórico-conceitual. *Perspectivas em Ciência da Informação*, v. 15, n. 2, p. 42-55, maio/ago. 2010.

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

WANG, J. *et al.* A boosted-trees method for name disambiguation. *Scientometrics,* v*.* 93, p. 391-411, 2012b.

WASSERMAN, S.; FAUST, K. *Social network analysis:* methods and applications. New York: Cambridge University Press, 1994.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. *Nature*, v. 393, p. 440-442, 1998.