



Revista de Administração Contemporânea

ISSN: 1415-6555

ISSN: 1982-7849

Associação Nacional dos Programas de Pós-graduação
em Administração

Araujo, Uajara Pessoa; Mendonça, Fabrício Molica de;
Campos, Rita de Cássia Leal; Silva, Lara Figueiredo e
Riscos e Possível Solução Associados às Amostras em Redes de Coautoria
Revista de Administração Contemporânea, vol.
22, núm. 5, 2018, Setembro-Outubro, pp. 762-785
Associação Nacional dos Programas de Pós-graduação em Administração

DOI: <https://doi.org/10.1590/1982-7849rac2018170340>

Disponível em: <https://www.redalyc.org/articulo.oa?id=84057863008>

- Como citar este artigo
- Número completo
- Mais informações do artigo
- Site da revista em redalyc.org

UAEM redalyc.org


Sistema de Informação Científica Redalyc
Rede de Revistas Científicas da América Latina e do Caribe, Espanha e Portugal
Sem fins lucrativos acadêmica projeto, desenvolvido no âmbito da iniciativa
acesso aberto

Artigos Tecnológicos:


Riscos e Possível Solução Associados às Amostras em Redes de Coautoria




Uajara Pessoa Araujo¹

 <https://orcid.org/0000-0001-5580-6587>


Fabício Molica de Mendonça²

 <https://orcid.org/0000-0001-8909-6843>

Rita de Cássia Leal Campos¹

 <https://orcid.org/0000-0001-6092-8810>

Lara Figueiredo e Silva¹

 <https://orcid.org/0000-0001-7334-8735>

Centro Federal de Educação Tecnológica de Minas Gerais, Departamento de Ciências Sociais Aplicadas, Belo Horizonte, MG, Brasil¹
Universidade Federal de São João del-Rei, Departamento de Ciências Administrativas e Contábeis, São João del-Rei, MG, Brasil²

Artigo recebido em 21.10.2017. Última versão recebida em 10.04.2018. Aprovado em 22.07.2018.
Editor Associado: Prof. Gustavo da Silva Motta.

Resumo

Já existe um conjunto razoável de trabalhos que aplicaram a sociometria e a teoria de redes para caracterizar o arranjo de pesquisadores e/ou de instituições de pesquisa subjacente a um objeto específico de interesse, seja ele Programa de Pós-Graduação, periódico, congresso, campo teórico ou técnico. Se o objeto de interesse é circunscrito, o censo é razoável e pode ser empregado. Caso contrário, trabalha-se com uma fração da população obtida por amostragem. Contudo, o uso de amostra tem riscos que, sendo ignorados, podem comprometer os achados. Frente a essa questão, este trabalho teve por objetivos avaliar os riscos decorrentes do emprego de amostras em estudos de redes de coautoria e propor um encaminhamento alternativo a simplesmente desconsiderá-los. Para tanto, foram feitas 300 simulações de uma rede de coautoria, reduzindo-a em 5, 10, 15, 20, 25 e 30%, para depois testar a extrapolação a partir do modelo linear. Os resultados indicam que mesmo amostras relativamente grandes podem ser enganadoras. Ainda assim, foi possível inferir algumas das características estruturais da população a partir do método em consideração, de tal forma que esse desenvolvimento pode vir a ser um recurso interessante a fim de conferir maior confiabilidade à pesquisa na área.

Palavras-chave: redes; redes de pesquisadores; redes de coautoria; sociometria; simulação.

Abstract

There is already a reasonable set of papers that applied sociometry and network theory to characterize the arrangement of researchers and/or research institutions underlying a specific object of interest, whether it be a Postgraduate Program, periodical, conference, or theoretical or technical field. If the object of interest is circumscribed, the census is reasonable and can be employed. Otherwise, you work with a fraction of the sample population. The use of such a sample has risks that, in theory, may compromise the findings. To meet this study's objectives – to evaluate the risks arising from the use of samples in co-authorship network analysis and to propose a more adequate approach than to simply disregard them – 300 simulations of a co-authorship network were made, reducing it by 5, 10, 15, 20, 25 and 30% to subsequently test the extrapolation from the linear model. Results indicate that even relatively large samples can be misleading. However, it was possible to infer some of the structural characteristics of the population from the method under consideration, in such a way that this development can turn out to be an interesting resource to confer greater reliability to the research in the area.

Keywords: networks; researchers' network; co-authorship network analysis; social network analysis (sociometry); simulation.

JEL codes: D85; L14; I23.

Introdução

Este *report* é um artigo tecnológico (Motta, 2017) destinado aos praticantes e interessados da sociometria, que pretende encaminhar um questionamento metodológico sobre o uso de amostragem em trabalhos de análise de redes, especificamente aqueles inscritos na cientometria. É comum a aplicação da sociometria com o intuito de descortinar as relações entre pesquisadores ou caracterizar um campo científico, em associação com discussões sobre a colaboração em ciências, a partir de trabalhos como os de Katz e Martin (1997), Melin (2000), Newman (2001, 2004), Barabási et al. (2002), Moody (2004), Rigby e Edler (2005) e Corley, Boardman e Bozeman (2006) – entre os mais citados.

Seria razoável admitir que a propagação de trabalhos com tal perfil também atingiu a academia brasileira, na medida em que um levantamento preliminar, feito em 30 de agosto de 2016, através do buscador Google Acadêmico, recuperou 273 trabalhos empíricos publicados no Brasil (dos quais, 134 em periódicos voltados à Administração) que empregaram a sociometria para investigar a produção acadêmica (aqui nomeados sócio-bibliométricos). Basicamente, essas pesquisas utilizam-se de coautorias e/ou citações para estabelecer a rede em estudo, traçar sociogramas, apurar características estruturais dos vértices e das redes e, algumas vezes, testar hipóteses da teoria de redes como homofilia, ligações preferenciais, **mundo pequeno** (*small world*) e centro-periferia.

Se o foco da pesquisa é bem delimitado, por exemplo: a evolução das ligações decorrentes da produção e dos projetos de pesquisa de um dado Programa de Pós-Graduação, é razoável esperar um tratamento censitário, que alcance 100% da produção e dos projetos.

Mas quando os eventos, indicadores de ligações de coautoria, não estão tão evidentes e nem mesmo totalmente conhecidos (por exemplo, toda a produção a respeito de **sustentabilidade** em anos recentes), é praxe pesquisar elegendo uma base de dados (ISI, *Scopus*, SciElo, entre as mais frequentes) ou um grupo de periódicos (por exemplo: aqueles da área, nos extratos mais elevados da classificação Qualis) ou anais de Congressos (como os EnANPAD) e adotar filtros adicionais para conter o número de eventos dentro das expectativas dos pesquisadores. O produto dessa seleção não deveria ser tomado como censo, nem tampouco, como uma amostra aleatória. Tratar-se-ia de uma amostra. Caso fosse empregado o método alternativo, seria obtida outra amostra diferente da primeira. Ao se fazer a apuração das grandezas estruturais da rede e dos vértices de cada amostra é razoável esperar divergências, eventualmente consideráveis.

Ademais, os pesquisadores podem ficar tentados a fazerem inferências a partir dessas amostras. Não necessariamente estatísticas, mesmo assim, inferências, do tipo: o campo científico em consideração não está ainda consolidado, pois a rede (o arranjo obtido da amostra de eventos) tem baixa densidade e é composta de múltiplos componentes.

No entanto, haveria riscos em tais procedimentos, tanto gerais – erros dos tipos I e II reconhecidos pela estatística (Hair, Black, Babin, Anderson, & Tatham, 2009) – quanto próprio da natureza sociométrica. Se a amostra fosse gerada aleatoriamente a partir de uma rede randômica com distribuição uniforme, a inferência seria apropriada, uma vez que o grau (ou *degree*: número de ligações distintas dos vértices) tem pequeno desvio-padrão – e, assim, a consequência da omissão de vértices e/ou de ligações pode ser antecipada de acordo com as características do arranjo (Borgatti, Carley, & Krackhardt, 2006). Mas esse não é o caso de uma rede empírica, ainda mais naquelas cujas distribuições do grau seguem a lei da potência, bem comum na sócio-bibliometria: pode ser que o vértice com maior grau não seja **capturado** na amostra; como consequência, haveria subestimação não quantificável do grau médio, uma das mais usuais medidas estruturais da rede. Não por outro motivo, Borgatti, Everett e Johnson (2013) apontam que, assim como outras falhas, tais “erros podem provocar **grandes impactos** nos indicadores de rede, particularmente, para algumas medidas de centralidade” (p. 37, grifo nosso), mesmo dramáticos (Kossinets, 2006).

Daí a motivação para este aprofundamento questionador, constituído como problema de pesquisa, da forma: quais são os riscos (específicos) de se empregar amostras em trabalhos empíricos sócio-

bibliométricos? O que, em complemento, ensejou o objetivo aplicado de propor/demonstrar um método simples, ainda que trabalhoso, para lidar com amostras.

A solução partiu do emprego de outro estudo sócio-bibliométrico, denominado PESQUISA BASE, que ofereceu a rede de coautoria de 573 trabalhos sociométricos publicados no Brasil até 2016. A partir desse conjunto, foram feitas 300 simulações, tal como indicado na seção de Métodos, que também detalha as técnicas estatísticas empregadas para a análise de dados. Os resultados e a análise estão na quarta seção, que precede as conclusões, capítulo que discute a validade dos achados das pesquisas sócio-bibliométricas que empregam amostras e do método proposto – completado pela obrigatória declaração das limitações do trabalho e pela proposição de outras pesquisas.

Antes, são apresentadas algumas breves considerações da literatura que serviram para a fundamentação da investigação e que sustentam o seu *design*.

Referencial Teórico

Coleta de dados em pesquisas sócio-bibliométricas baseadas em coautoria

É recorrente o uso de repositórios internacionais para recuperar artigos de periódicos de alto impacto e daí extrair redes de coautoria para revelar a colaboração entre cientistas, instituições e países (Stefano, Giordano, & Vitale, 2011). Entre os repositórios mais utilizados estão o *Web of Science* (também conhecido como ISI Web of Knowledge), o MEDLINE, o *Scopus*, o *ScienceDirect* (no Brasil, o equivalente, é o *Scientific Electronic Library Online* [SciELO]). A uso desses repositórios pode ser explicado pelo custo mais razoável, menor carga de esforço e maior acurácia quando confrontados com meios alternativos (Stefano, Fuccella, Vitale, & Zaccarin, 2013).

No entanto, a escolha por um ou por outro repositório, ou outro mecanismo de recuperação, tem impacto significativo nos achados da pesquisa. Essa constatação é apresentada no trabalho de Stefano, Fuccella, Vitale e Zaccarin (2013), voltados para estudar a rede de cientistas do campo da estatística na Itália. Apurando as redes de cada uma das três bases empregadas, os autores depararam-se com diferenças estruturais significativas, como, por exemplo, na indicação (e não indicação) do fenômeno do **mundo pequeno**, dependendo da base.

Mas esse não é o único desafio da pesquisa sócio-bibliométrica baseada na coautoria. Ao se escolher um repositório de artigos, por mais completo, exclui-se da população outras formas de comunicação, como livros, anais em Congressos, publicação em revistas comerciais e jornais – e não há uma justificativa plausível para essa estratégia, sendo assim recomendável o uso combinado de todas as formas de divulgação científica (Stefano et al., 2011) e outros indícios de colaboração, como projetos de pesquisas em comum. O uso de múltiplas fontes de dados já era antevisto como interessante e mesmo necessário desde Hicks (1998), voltado para identificar o grau de dificuldade de atingir a cobertura plena da literatura, em especial para as ciências sociais, e o impacto disso na bibliometria.

Cabe o registro que a opção por uma amostra em detrimento da população (mesmo que não proposital, mas consequência do método de coleta de dados) depara-se com uma respeitada advertência sociométrica: a depender das características estruturais da rede, um subconjunto dela derivado, mesmo que aleatório, pode não guardar tais propriedades e assim, a extrapolação desse subconjunto não é razoável (Stumpf, Wiuf, & May, 2005).

Erros na pesquisa sociométrica

A sociometria depara-se com desafios próprios. Apenas para introduzir o ponto: se é válido o achado sociométrico mais popular, dos seis passos médios de separação entre as pessoas (Milgram, 1967), implicando conectividade potencialmente planetária e entendendo redes como depositárias de

ativos que podem ser acionados através de conexões (familiares, afetivas, profissionais ou mesmo mero conhecimento) diretas ou indiretas, fortes ou fracas (Burt, 1980; Granovetter, 1985; Lin, 1999; Sewell, 1992), então qualquer delimitação de rede de pessoas parece trazer consigo arbitrariedades redutoras – o que é reconhecido como *boundary specification problem* (BSP).

O BSP está relacionado com a definição de regras para a inclusão de atores e de ligações na rede sob investigação (Kossinets, 2006; Laumann, Marsden, & Prensky, 1983). Kossinets (2006) não se permite dúvida: BSP é o maior problema epistemológico da pesquisa sociométrica, com três soluções encontráveis na literatura: (a) a nominal: alguma definição formal e arbitrária para inclusão de vértices, incluindo seus atributos e os tipos de ligações e de eventos; (b) a realista-cognitivista: os próprios atores definem os limites, uma vez que a rede seria tomada como fato social somente se os atores estiverem conscientes dela (consciência subjetiva compartilhada coletivamente); e (c), a empírica, identificando-se quem interage com quem e em qual contexto.

Além da BSP, a sociometria tem avançado e procura sistematicamente lidar com outras de suas dificuldades intrínsecas. Nesse sentido, por exemplo, Robins, Pattison e Woolcock (2004) propuseram-se a estudar a situação de algumas pessoas que, dentro da população da pesquisa, não responderem ao questionário sociométrico (que busca captar ligações), mas foram citadas como parceiros pelos respondentes. Robins et al. (2004) desenvolveram e aplicaram seu modelo grafo aleatório-exponencial (p^*) em duas possibilidades: distribuição aleatória ou não aleatória de não respondentes, e concluíram que, no primeiro caso, ainda é possível ter resultados consistentes sem a necessidade de acrescentar um modelo heurístico complementar quando se lida com não respondentes não distribuídos aleatoriamente.

Erros na identificação dos vértices são potencialmente originados pela presença de homônimos, grafias diferentes, mudanças de nomes, nomes incompletos, abreviações e erros operacionais. Eles podem ocasionar quatro dos seis erros na classificação de Wang, Shi, McFarland e Leskovec (2012): ausência, inclusão, agregação e desagregação indevidas (os outros erros seriam relativos à ausência e inclusões indevidas de ligações). Tais erros seriam críticos na sócio-bibliometria, também para Barbastefano, Souza, Costa e Teixeira (2013, 2015), que estudaram a estabilidade de algumas características de rede (densidade, grau médio, distribuição de graus, componente maior, distância média, diâmetro e coeficiente de clusterização) e dos vértices (centralidades de grau, por proximidade e por intermediação), permitindo, propositalmente, erros de identificação (ao declinar do tratamento para redução das ambiguidades) de coautores em uma dada rede de pesquisa em sustentabilidade, gerando-a através de: nome completo, nome abreviado e sobrenome + inicial do primeiro nome. Os resultados ratificam o estudo de Borgatti et al. (2013): foi encontrado comprometimento significativo nas medidas.

Para Wang et al. (2012), em trabalho de natureza similar, a intensidade do comprometimento depende das próprias características da rede; por exemplo, seu coeficiente de clusterização. Nessa sua pesquisa, ao mesmo tempo em que destacam o avanço de algumas técnicas computacionais para a **limpeza** dos dados, tais autores admitem que o desafio da confiabilidade da identificação estará ainda pendente mesmo se passar a ser aceita uma identificação única para os pesquisadores. Afinal, os dados do passado, frequentemente usados em pesquisas sociométricas, não seriam tão facilmente atualizados para a nova forma de registro (Barbastefano, Souza, Costa, & Teixeira, 2013).

Um estudo de Costenbader e Valente (2003) é particularmente interessante ao testar a estabilidade de 11 medidas de centralidade **reduzindo** oito arranjos empíricos (de campos diversos, com restrições e cenários distintos, cada um tomado com uma **população**) de 10 em 10%. Os autores anotaram que algumas medidas são mais estáveis que outras, mas essa estabilidade dependeria das propriedades da rede. E, ao encontrar coeficientes de correlação relativamente elevados entre centralidade de grau e centralidade *eigenvector* da **população** e centralidade de grau e centralidade *eigenvector* das subredes simuladas, Costenbader e Valente (2003) propuseram que a pesquisa sociométrica ainda poderia oferecer alguma contribuição ao entendimento de fenômenos mesmo que o pesquisador não tenha conseguido coletar dados de todos os membros da população.

De forma similar, mas interessados em medidas para a rede e não nas características de seus vértices, Smith e Moody (2013) lidaram com 12 redes empíricas, examinando o efeito da exclusão de

vértices nas características de centralidade, centralização, topologia e homofilia. Eles encontraram, por exemplo, que a centralidade por aproximação é mais sensível à perda de dados do que a centralidade por proximidade; e que redes maiores e mais centralizadas apresentam maior robustez à perda de vértices – sinalizando que a perda de dados pode ter efeito, considerável ou insignificativo, dependendo da rede e do que está se medindo.

Do exposto, percebe-se que um modo de lidar com erros é estudá-los. Kossinets (2006) sintetiza o *design* – convencional e preferido inclusive na literatura consultada de pesquisas com esse propósito, que consiste em escolher uma rede grande (ou construí-la através de grafos aleatórios), assumir que a rede é completa, remover uma fração de entidades para simular fontes diferentes de erros, medir as propriedades da simulação e comparar com as medidas **verdadeiras**. Tal estratégia foi adotada e aplicada nesta pesquisa com algumas particularidades, como apresentado a seguir.

Método

A pesquisa tem propósitos descritivo e prescritivo, em sua natureza objetivista, sustentado por análise quantitativa: basicamente, o método consistiu em simular diversas redes a partir de uma rede inicial e analisar o comportamento das grandezas estruturais obtidas.

Da PESQUISA BASE, com a sua coleção de 573 trabalhos, colheu-se o arranjo de coautoria que aqui é tratada como **origem** (S00). Para efeito do estudo, S00 foi recepcionado como o universo. No caso, um universo com 1013 vértices e 1531 ligações diáticas bidirecionais (*edges*) distintas, distribuídos em 218 componentes, o maior deles com 87 vértices.

A mesma rede foi estudada por Araújo et al. (2017) com outros propósitos, quando se revelou que S000 é (a) de baixa densidade, em torno de 0,0030; e (b), aderente aos modelos: do mundo pequeno, de centro-periferia, de homofilia e da escolha preferencial. Tais indicações robusteceram a conveniência de emprego de S00 na presente investigação: ela teria características próximas a algumas das redes sócio-bibliométricas encontradas na PESQUISA BASE, como por exemplo, nos trabalhos de Rossoni e Guarido (2007) e Rossoni (2014).

Da S00, foram extraídos seis subconjuntos (nomeados **classes**), cada um com 50 redes, obtidas excluindo-se aleatoriamente (e com reposição entre os conjuntos) uma fração dos trabalhos geradores da S00, conforme Tabela 1. Nela, por exemplo, o subconjunto de 50 redes da classe S30 foi obtido trabalhando-se com amostras de 401 itens, retidos de 50 seleções aleatórias a partir dos 573 trabalhos originais.

Tabela 1

Amostragem para as Simulações

Rede e classes	Fração excluída (%)	Trabalhos (n)
S00	0	573
S05	5	544
S10	10	516
S15	15	487
S20	20	458
S25	25	430
S30	30	401

As características estruturais de cada uma das 300 redes, bem como de seus vértices, foram determinadas com a utilização dos *softwares* Pajek e Ucinet. Para as redes, foram apurados diversos indicadores. Por economia, são destacadas apenas as características mais usuais: Número de vértices (V), Número de ligações (L), Grau ou *Degree* médio (GM), Densidade (d), Número de componentes (NC), Número de vértices do maior componente (VMC), Diâmetro (ϕ), Conectividade (Conec) e Distância Média (DM). Para os vértices, o foco ficou restrito às medidas de centralidade: de grau (*degree*); por intermediação (*betweenness centrality*); e por proximidade (*closeness centrality*); e grau de redundância (*aggregate constraint*).

Todas as medidas foram transferidas para o pacote IBM SPSS. O SPSS foi encarregado de:

- apurar as estatísticas descritivas das características estruturais de cada conjunto (incluindo média e mediana, desvio-padrão, maior e menor valor, amplitude, intervalo de 95% de confiança da média, assimetria e curtose);
- desenhar os *box-plots* e os histogramas;
- processar os testes de normalidade para cada variável de cada subconjunto;
- e, quando recomendado, proceder os testes de regressão (no caso, empregando-se o modelo linear, por simplicidade), com a determinação de curvas e estimação de erros.

Cabe uma observação inicial: a regressão, como outras técnicas estatísticas, tem como requisito a normalidade dos dados (Hair et al., 2009). Ainda que alguns trabalhos dispensem essa exigência, aqui – como será visto – não haveria espaço para tal complacência. Assim, as regressões foram precedidas pela devida comparação da distribuição de frequência dos dados frente à distribuição de Gauss. Entre os testes de normalidade, optou-se pelo método Shapiro-Wilk, aceitando-se como limite o nível de significância de 0,050 e admitindo-se como hipótese nula $H_0 \sim N$ (trata-se de uma distribuição normal). Ressalta-se que, mesmo para 50 dados, o teste escolhido já se revela com poder razoável, de rejeitar corretamente a hipótese H_0 de normalidade (Torman, Coster, & Riboldi, 2012).

Resultados e Análise

As indicações colhidas dos seis conjuntos de simulação estão apresentadas e discutidas nas duas seções seguintes. Na primeira, são tratados os parâmetros de rede, trazendo a estatística descritiva, os testes de normalidade e, por fim, o resultado de testes de regressão. A análise voltada para os vértices é o objeto da segunda parte do capítulo.

Características estruturais da rede

Estatística descritiva

A Tabela 2 exibe a estatística descritiva das medidas de rede mais comuns encontradas na sócio-bibliometria estudada. As colunas apresentam sequencialmente: a característica estrutural em questão; a média e o desvio-padrão das 50 simulações; seu coeficiente de variação, em %; o maior e o menor valor encontrados; a amplitude dessa variação; o valor para a rede de origem; e, finalmente, a comparação absoluta e a relativa (em %) entre a média das simulações e o valor de origem. Nas linhas, estão as indicações dos seis conjuntos de simulações (classes S05 a S30).

Tabela 2

Características Estruturais das Redes

Item	Rede	Média (A)	Desvio (B)	Cv/100 (B)/(A)	Maior (D)	Menor (E)	R (D) - (E)	S00	Δ (A) - S00	$\delta/100$ $\Delta/S00$
V	S05	972,7	6,2	0,6	987	958	29	1013	-40,3	-4,0
	S10	938,2	8,2	0,9	951	918	33	1013	-74,8	-7,4
	S15	895,7	11,7	1,3	924	868	56	1013	-117,3	-11,6
	S20	853,9	12,1	1,4	881	831	50	1013	-159,1	-15,7
	S25	813,3	12,1	1,5	839	789	50	1013	-199,7	-19,7
	S30	765,4	14,6	1,9	798	737	61	1013	-247,6	-24,4
L	S05	1459,6	16,2	1,1	1494	1416	78	1531	-71,4	-4,7
	S10	1399,8	20,0	1,4	1441	1351	90	1531	-131,2	-8,6
	S15	1330,3	23,2	1,7	1375	1264	111	1531	-200,7	-13,1
	S20	1261,2	24,5	1,9	1317	1200	117	1531	-269,8	-17,6
	S25	1188,7	32,6	2,7	1242	1122	120	1531	-342,3	-22,4
	S30	1110,3	36,4	3,3	1187	1035	152	1531	-420,7	-27,5
GM	S05	3,001	0,021	0,7	3,037	2,932	0,105	3,023	-0,022	-0,7
	S10	2,984	0,027	0,9	3,040	2,905	0,135	3,023	-0,039	-1,3
	S15	2,970	0,028	0,9	3,025	2,909	0,115	3,023	-0,053	-1,7
	S20	2,954	0,043	1,5	3,034	2,824	0,211	3,023	-0,069	-2,3
	S25	2,921	0,053	1,8	3,028	2,795	0,234	3,023	-0,102	-3,4
	S30	2,900	0,059	2,0	2,997	2,734	0,263	3,023	-0,123	-4,1
d	S05	0,0031	0,0000	0,8	0,0031	0,0030	0,0001	0,0030	0,0001	3,4
	S10	0,0032	0,0000	1,1	0,0033	0,0031	0,0002	0,0030	0,0002	6,6
	S15	0,0033	0,0000	1,5	0,0035	0,0032	0,0002	0,0030	0,0003	11,2
	S20	0,0035	0,0001	2,1	0,0036	0,0033	0,0003	0,0030	0,0005	16,0
	S25	0,0036	0,0001	1,9	0,0037	0,0034	0,0003	0,0030	0,0006	20,4
	S30	0,0038	0,0001	2,2	0,0040	0,0036	0,0004	0,0030	0,0008	27,0
ϕ	S05	9,8	0,6	5,9	10	7	3	10	-0,2	-1,6
	S10	9,5	0,9	9,3	10	7	3	10	-0,5	-5,0
	S15	9,1	1,3	13,9	10	7	3	10	-0,9	-9,2
	S20	8,9	1,2	13,7	10	6	4	10	-1,1	-10,8
	S25	8,1	1,5	18,9	10	6	4	10	-1,9	-19,0
	S30	7,8	1,4	18,3	10	6	4	10	-2,2	-21,6
NC	S05	212,7	2,4	1,1	218	207	11	218	-5,3	-2,4
	S10	208,1	3,6	1,7	216	200	16	218	-9,9	-4,5
	S15	201,4	4,0	2,0	210	190	20	218	-16,6	-7,6
	S20	194,8	5,7	2,9	205	184	21	218	-23,2	-10,6
	S25	189,8	5,6	3,0	203	174	29	218	-28,2	-13,0
	S30	181,4	6,3	3,5	201	169	32	218	-36,6	-16,8

Continua

Tabela 2 (continuação)

Item	Rede	Média (A)	Desvio (B)	Cv/100 (B)/(A)	Maior (D)	Menor (E)	R (D) - (E)	S00	Δ (A) - S00	$\delta/100$ $\Delta/S00$
VMC	S05	79,7	9,5	11,9	87	44	43	87	-7,3	-8,4
	S10	73,1	14,2	19,4	87	43	44	87	-13,9	-16,0
	S15	66,1	16,8	25,4	87	36	51	87	-20,9	-24,1
	S20	61,8	14,9	24,2	82	39	43	87	-25,2	-29,0
	S25	51,9	15,7	30,3	79	29	50	87	-35,1	-40,3
	S30	47,1	14,0	29,8	75	24	51	87	-39,9	-45,9

Como esperado e de acordo com a Tabela 2, ao se excluir uma fração dos trabalhos, caem os números de autores e das ligações estabelecidas entre os vértices remanescentes, bem como são eliminados aqueles laços conectores de vértices excluídos e, desses com os remanescentes. Também são reduzidos: o grau médio, o diâmetro, o número de componentes, assim como o número de vértices do maior componente da rede. Acrescente-se que se $GM = (2 \times L) / V$ em um grafo bidirecional (constituído com *edges*) como naqueles resultantes de coautorias, portanto, o decréscimo de GM implica maior redução percentual do número de vértices quando comparada a essa redução no número de ligações.

Todavia, a densidade aumenta com a redução do tamanho da amostra. Isso é esperado matematicamente, pois, se a densidade segue a fórmula $d = (2 \times L) / [V \times (V - 1)]$, seria necessária uma diferença dos decréscimos relativos ainda maior do que a observada para haver redução da densidade com a redução da amostra.

Sendo assim, pelo menos para o universo considerado, trabalhar com amostra subestima o número de vértices, de ligações e de componentes; e também o grau médio, o diâmetro e o tamanho do maior componente; por outro lado, superestima a densidade.

Isso suscita a questão de estabelecer a melhor forma de combinar V e L, quer através do grau médio ou pela densidade. Pela literatura, a densidade não deve ser utilizada na comparação de redes de tamanhos consideravelmente desiguais (Borgatti et al., 2006). Pelo presente estudo, (a) o comportamento do grau médio é mais facilmente entendível; (b) o grau médio teve sistematicamente menor Coeficiente de Variação dentro de cada classe; e (c) ele é bem mais estável quando comparado com a S00 – a maior diferença entre grau médio da classe frente a origem foi de 4,1% mesmo quando a amostra foi 70% da população frente a 27,0 % da densidade.

Seguindo, é notável a sobreposição parcial com a classe anterior: mesmo com uma base menor, eventualmente e por exemplo, uma rede da classe S10 pode ter algumas das características de outra rede, essa da classe S05 – isso em função das amplitudes apuradas, de valor bem significativo.

Registra-se ainda que foi encontrado um nível de significância bicaudal de 0,000 para as correlações (tomadas em pares, de todas as 300 simulações) das grandezas V, L, GM, d, NC, ϕ , VMC, Conec e DM e, desses com o número de artigos tratados em cada amostra (NA). A Tabela 3 apresenta o Coeficiente de Pearson para as correlações, com destaque em negrito para aquelas mais fortes, que implicam um R^2 superior a 0,75. É interessante reparar que os dados indicam uma correlação de intensidade apenas média entre GM com V e L (R de 0,668 e 0,762, respectivamente) enquanto d foi encontrada em correlação forte tanto com V quanto com L.

Tabela 3

Coefficiente de Pearson (R) (sig. bicaudal = 0,000)

	L	CNC	GM	d	ϕ	VMC	Conec	DM	NA
V	0,991	0,925	0,668	-0,973	0,507	0,618	0,453	0,557	0,987
L		0,875	0,762	-0,938	0,520	0,633	0,496	0,573	0,976
CNC			0,394	-0,958	0,383	0,485	0,246	0,420	0,911
GM				-0,511	0,443	0,532	0,570	0,499	0,646
d					-0,472	-0,574	-0,378	-0,515	-0,961
ϕ						0,912	0,846	0,940	0,506
VMC							0,918	0,977	0,618
Conec								0,943	0,465
DM									0,558

Contudo, como previamente sinalizado e antes de partir para a construção de modelos, é necessária uma investigação adicional, analisando-se as distribuições das grandezas através do teste de normalidade.

Teste de normalidade

Os resultados do teste de normalidade para as principais características estruturais da rede estão apresentados na Tabela 4.

Tabela 4

P-value Teste Shapiro-Wicks

	Amostra					
	S05	S10	S15	S20	S25	S30
V	0,501	0,250	0,239	0,433	0,833	0,255
L	0,191	0,612	0,586	0,923	0,164	0,791
GM	0,023	0,798	0,785	0,162	0,244	0,254
d	0,000	0,000	0,000	0,000	0,000	0,000
ϕ	0,000	0,000	0,000	0,000	0,000	0,000
NC	0,201	0,705	0,700	0,234	0,249	0,170
VMC	0,000	0,000	0,000	0,000	0,000	0,000

Na Tabela 4, estão destacados os casos de $P\text{-value} \leq 0,050$; portanto, implicando a rejeição da hipótese nula, de normalidade. Em consequência, é razoável assumir que, para o caso e as amostras considerados, as distribuições da densidade, do diâmetro e da quantidade de vértice do maior componente do arranjo não se aproximam da normal. Isso em oposição ao comportamento das distribuições do número de vértices, do número de linhas e do número de componentes, para as quais a hipótese nula sobreviveu.

Já quanto ao grau médio, esse falhou apenas no teste correspondente à redução de 5% do número de trabalhos considerados – mas isso foi considerado suficiente para excluí-lo da aplicação da regressão. Esse resultado (falha no teste de normalidade para S05 e positivo para as demais classes) é inesperado,

e para a qual não foi encontrada uma explicação outra que aquela derivada da aleatoriedade no processo de construção das simulações.

Felizmente, V e L são as grandezas mais básicas e independentes entre si, das quais derivam outras características, como d e GM; enquanto o NC supre uma imagem complementar e interessante sobre a fragmentação do arranjo.

Entretanto, apenas como exemplo e considerando a relevância do maior componente – usualmente empregado para o teste da hipótese de mundo pequeno, é apresentada na Figura 1 a distribuição do número de vértices do maior componente para a amostra S15.

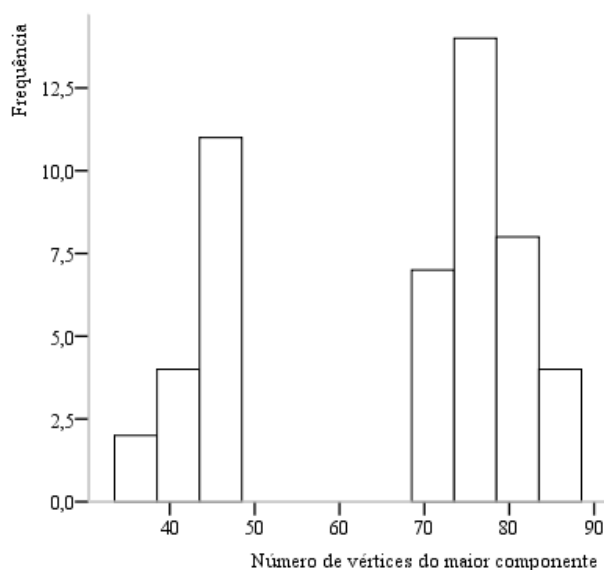


Figura 1. Distribuição de Frequência de VMC para a Amostra S15

Da Figura 1 e dos demais histogramas da pesquisa (não apresentados, por economia), propõe-se que o desvio da normalidade não deve ser ignorado: no destaque, trabalhando-se uma amostra de 85% do total de itens utilizados para S00, ora pode-se afirmar que o componente pode ser tão pequeno a ponto de não ter 40 vértices ou, ora, tão grande a ter mais de 80 vértices (como antecipado pelo elevado coeficiente de dispersão – $Cv/100$, na Tabela 2), evidenciando o que na Figura 1 poderia ser considerado como dois conjuntos distintos de dados. Esse fenômeno, se não é surpreendente frente à literatura consultada, é expressivo devido à sua extensão e às consequências derivadas da constatação de que a incerteza nas medidas sociométricas indicativas da coesão da rede pode ser considerável.

Regressão linear

As Figuras 2A, 2B e 2C trazem os *box-plots* das três características que sobreviveram ao teste de normalidade: V, L e NC, em função do número de artigos considerados em cada uma das seis amostras.

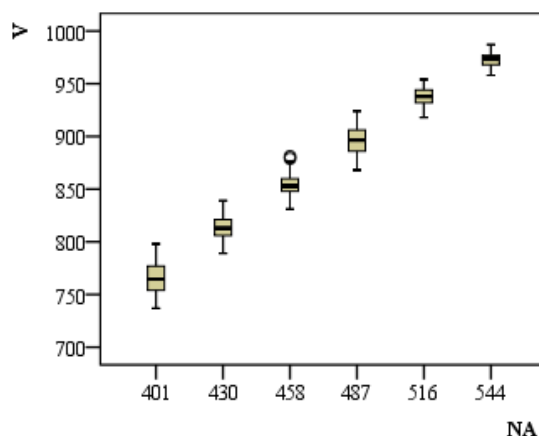


Figura 2A. Vértices e Número de Artigos

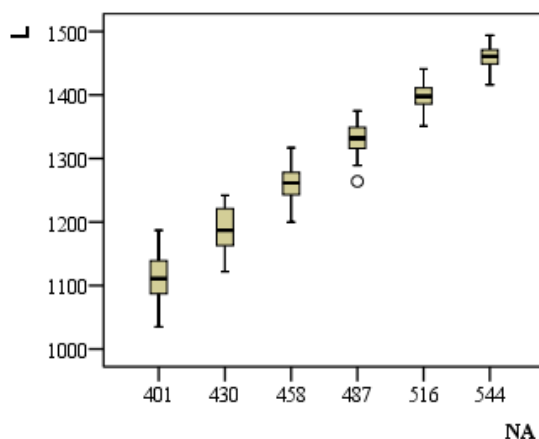


Figura 2B. Linhas e Número de Artigos

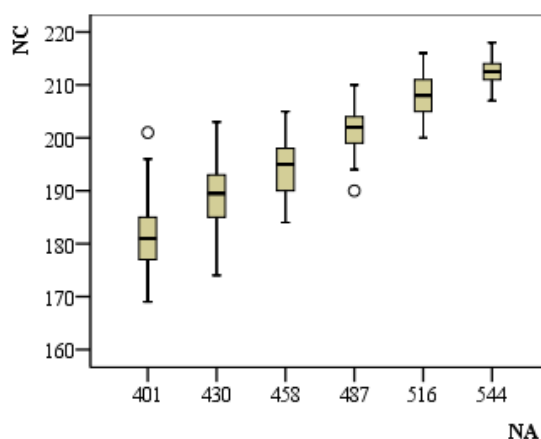


Figura 2C. Número de Componentes e Número de Artigos

As Figuras 2A, 2B e 2C apontam para uma provável viabilidade do modelo linear. Assim, tomando como variável independente o número de artigos empregados nas amostras das seis classes e como variável dependente o número de vértices, de ligações e de componentes, são obtidos os resultados apresentados nas primeiras colunas da Tabela 5, que ainda traz, nas suas duas últimas colunas, os valores calculados (usando-se a equação da reta) e os valores **reais**, quando o número de artigos for igual a 573 (aquele relativo ao ponto considerado como a **população** nesta pesquisa).

Tabela 5

Regressão Linear, Tal Que $\hat{y} = B + Ax$; Onde x = Número de Artigos Para a Amostra

	y	R ² ajust.	E. P. da estimativa	Sig.	B			A			S00	
					Valor	E. P.	Sig.	Valor	E. P.	Sig.	Cálc.	Real
S30, S25, S20	V	0,886	13,011	0,000	143,816	19,643	0,000	1,553	0,046	0,000	1034	1013
	L	0,794	31,477	0,000	48,812	47,522	0,306	2,648	0,110	0,000	1566	1531
	NC	0,463	5,913	0,000	87,191	8,927	0,000	0,236	0,021	0,000	222	218
S30, S25, S20, S15	V	0,935	12,727	0,000	162,508	12,528	0,000	1,508	0,028	0,000	1027	1013
	L	0,885	29,647	0,000	85,521	29,184	0,004	2,561	0,066	0,000	1553	1531
	NC	0,639	5,490	0,000	90,536	5,404	0,000	0,228	0,012	0,000	221	218
S30, S25, S20, S15, S10	V	0,963	11,988	0,000	169,942	8,596	0,000	1,491	0,019	0,000	1024	1013
	L	0,930	31,477	0,000	107,299	20,123	0,000	2,510	0,044	0,000	1546	1531
	NC	0,763	5,154	0,000	90,902	3,696	0,000	0,227	0,008	0,000	221	218
S30, S25, S20, S15, S10, S05	V	0,975	11,455	0,000	187,814	6,428	0,000	1,450	0,014	0,000	1019	1013
	L	0,953	26,700	0,000	136,366	14,982	0,000	2,444	0,032	0,000	1537	1531
	NC	0,830	4,832	0,000	95,036	2,711	0,000	0,218	0,006	0,000	220	218

A partir da Tabela 5, é razoável admitir que o modelo escolhido apresentou um ajuste considerável (R^2 ajustado de médio a elevado, com nível de significância de 0,000), com coeficientes razoavelmente ajustados, mesmo para a amostra S30, S25 e S20 – portanto, com apenas 150 simulações (nesse caso, com a exceção de NC, com R^2 ajustado inferior a 0,5; aparentemente, o NC requer maior número de simulações para avanço em confiabilidade, quando comparado a V e L).

Mais relevante ainda é indicar que foi apurada uma diferença pouco considerável em termos relativos entre o valor real e o valor calculado para V, L e NC usando-se o modelo, mesmo **extrapolando** os dados obtidos do conjunto de amostras decorrentes da redução de pelo menos 20% da quantidade de artigos que deram origem ao arranjo original. Isso também implicou uma expressiva concordância quando se apuram GM e d através de V e L calculados nos quatro conjuntos da Tabela 5, pois $GM_{\text{Calculado}}$, (obtido matematicamente, e não por regressão) varia de 3,017 a 3,030, para $GM_{\text{Real}} = 3,023$; e a $d_{\text{Calculada}}$ (também obtida matematicamente, e não por regressão) varia entre 0,0029 e 0,0030, para $d_{\text{Real}} = 0,0030$.

Avançando ainda mais na extrapolação (agora sem comprovação, mas apenas com base na dinâmica observada) – o que implica a aceitação de maior faixa de tolerância e/ou menor grau de confiança da estimativa, acarreta que se fossem trabalhados 30% a mais de artigos (adição estimada como provável para alcançar toda a população da PESQUISA BASE), portanto $NA = 745$, os valores de V, L e NC estariam na vizinhança de 1268, 1957 e 257, respectivamente, implicando uma densidade de meros 0,0024 e um Grau Médio próximo a 3,087 – obtidos das expressões matemáticas destas grandezas: função de V e de L.

Contudo, não foi encontrada uma solução aceitável para inferir o comportamento de outras grandezas da rede, como o número de vértices no componente maior e o diâmetro. Tampouco, antecipasse, foi encontrada uma regra para estimar as principais medidas de centralidade da população a partir de uma amostra, dado o comportamento destacado na segunda parte da análise.

Centralidade dos vértices proeminentes

Sendo impraticáveis a análise mais detida e a correspondente apresentação das características de 1013 vértices, foi estabelecido um corte arbitrário: grau igual ou superior a sete na S00. Isso resultou no destaque de um bloco de 44 autores (aqui, **Proeminentes**), cabendo a ressalva de que os dados apurados correspondem ao conjunto completo, de todos os vértices presentes em cada simulação.

Também por facilitação e sem comprometimento da discussão, das seis classes apuradas, a apresentação ficou restrita às redes S05 e S30, tal como nas Figuras 3A a 6B, que têm, na abcissa, os códigos dos vértices (observando-se que: nos casos de grau, da centralidade por proximidade e da centralidade por intermediação, tais vértices foram sequenciados a partir do maior valor encontrado para a grandeza em consideração; enquanto que, no caso do grau de redundância, eles foram sequenciados a partir do menor valor); e, no eixo das ordenadas, as características estruturais em relevo.

Nas Figuras 3A e 3B, as linhas apresentam o maior, o menor e a média dos valores de grau para cada vértice, encontrados entre as 50 simulações da classe indicada. Como o maior valor corresponde ao grau da S00 (com uma única exceção de um vértice na Figura 3B), a linha da S00 não é apresentada, para evitar superposição. Já nas Figuras 4A, 4B, 5A, 5B, 6A e 6B, com a mesma estruturação das Figuras 3A e 3B, as linhas com os valores de S00 também estão incluídas.

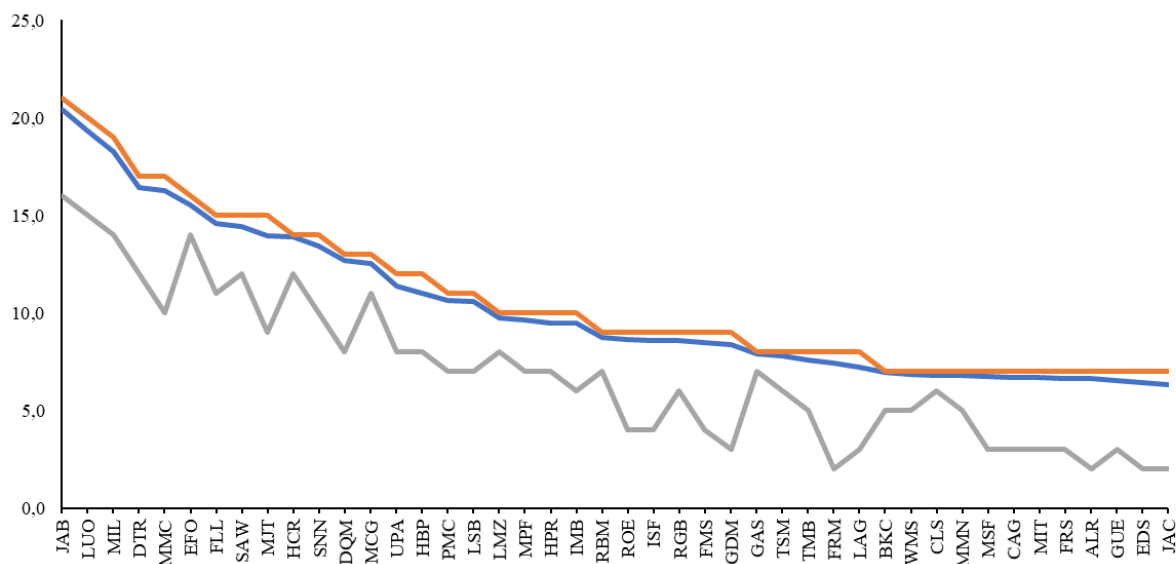


Figura 3A. Grau, Classe S05 (50 Simulações)

Legenda: cinza = menor; azul = média; laranja = maior.

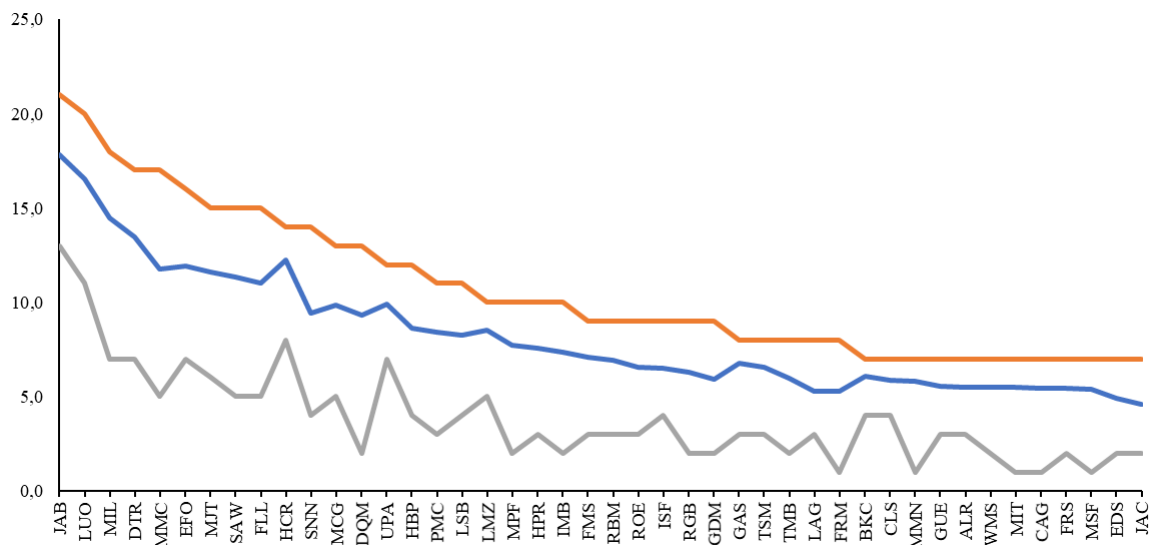


Figura 3B. Grau, classe S30 (50 simulações)

Legenda: cinza = menor; azul = média; laranja = maior.

As Figuras 3A e 3B (complementadas pelas estatísticas descritivas de S10, S15, S20 e S25, não apresentadas, mas disponíveis por solicitação) indicam que, de S05 a S30, a amplitude (Maior – Menor) aumentou e a média afastou-se do maior valor (e assim, de S00), aproximando-se da mediana. Mas as linhas apresentam um comportamento irregular (picos e declives), não obedecendo a qualquer padrão percebido.

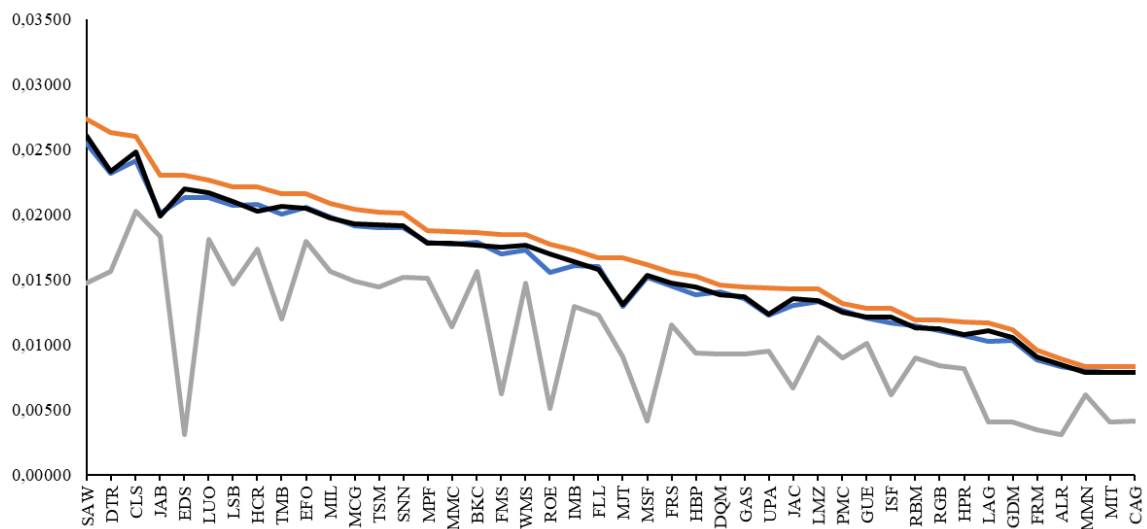


Figura 4A. Centralidade por proximidade, classe S05 (50 simulações)

Legenda: cinza = menor; azul = média; preto: s00; laranja = maior.

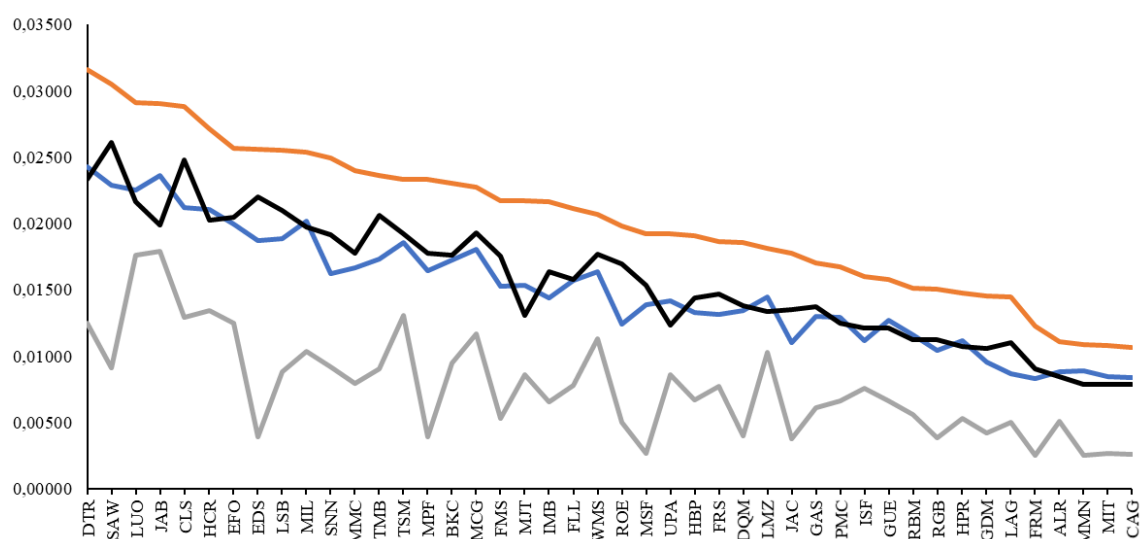


Figura 4B. Centralidade por proximidade, classe S30 (50 simulações)

Legenda: cinza = menor; azul = média; preto: s00; laranja = maior.

Nas Figuras 4A e 4B, quem se aproxima de S00 é a média, com os maiores valores bem superiores ao arranjo de origem. Todavia, semelhante ao grau, também na centralidade por proximidade, a amplitude é crescente com a diminuição do número de artigos que deram origem às redes. E, em complemento, a imprevisibilidade do comportamento das linhas se mostra mais evidente e verificam-se trocas de posição quanto ao maior valor obtido.

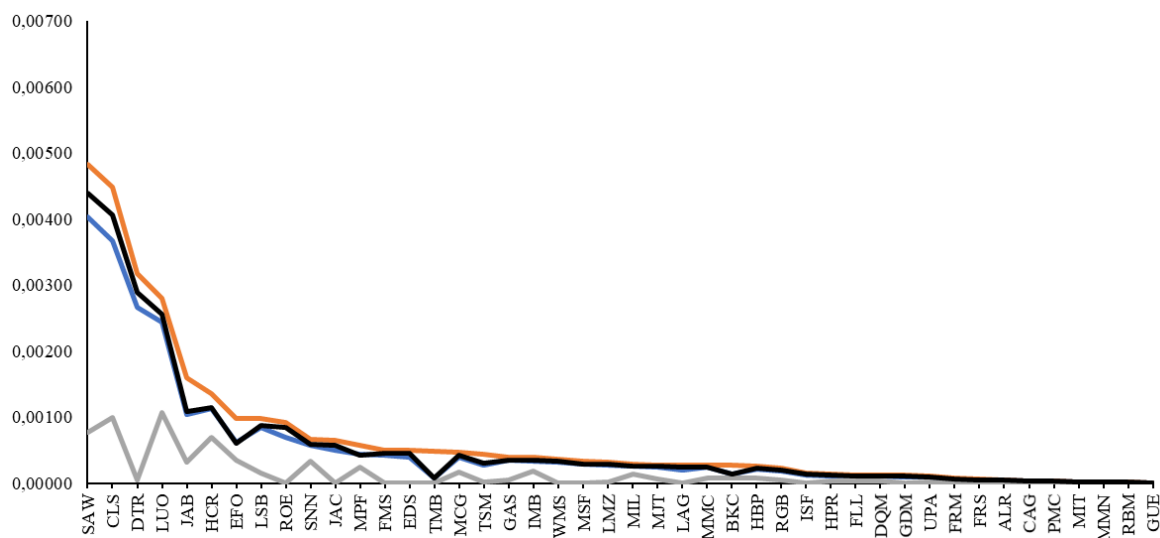


Figura 5A. Centralidade por intermediação, classe S05 (50 simulações)

Legenda: cinza = menor; azul = média; preto: s00; laranja = maior.

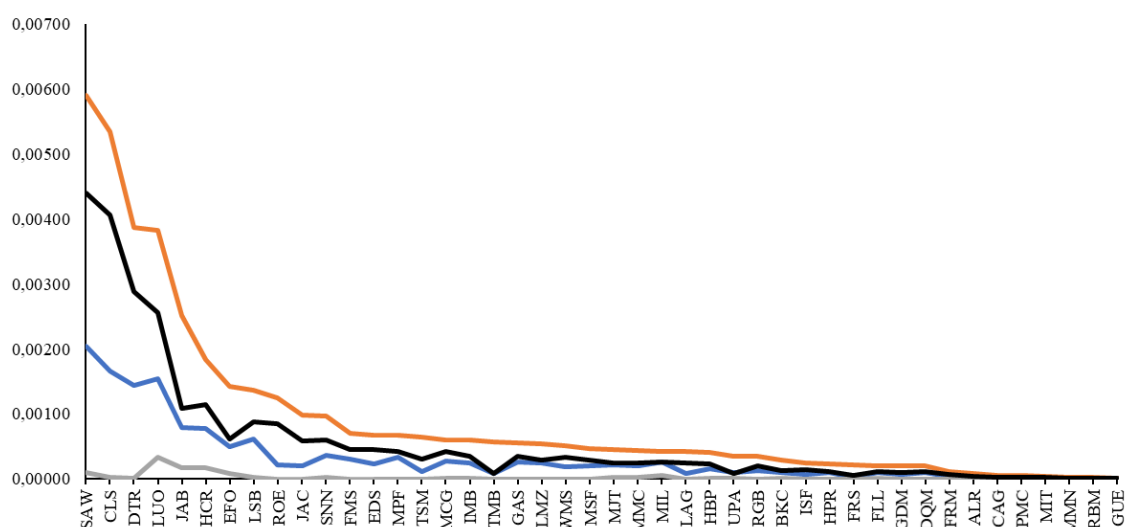


Figura 5B. Centralidade por intermediação, classe S30 (50 simulações)

Legenda: cinza = menor; azul = média; preto: s00; laranja = maior.

As observações relativas à centralidade por intermediação (Figuras 5A e 5B) são as mesmas daquelas relativas às indicações da centralidade por proximidade, exceto quanto a troca de posições (a ordem se mantém para os principais vértices entre os proeminentes). Destaca-se, ainda, que os valores de S00 são sistematicamente maiores ou pelo menos iguais aos valores médios (a média da centralidade por intermediação subestimaria o valor correspondente na população). E, apenas como nota, registra-se que a grandeza em consideração é altamente discriminatória (confere a maioria absoluta dos vértices valores próximos ou iguais a zero, destacando apenas poucos vértices).

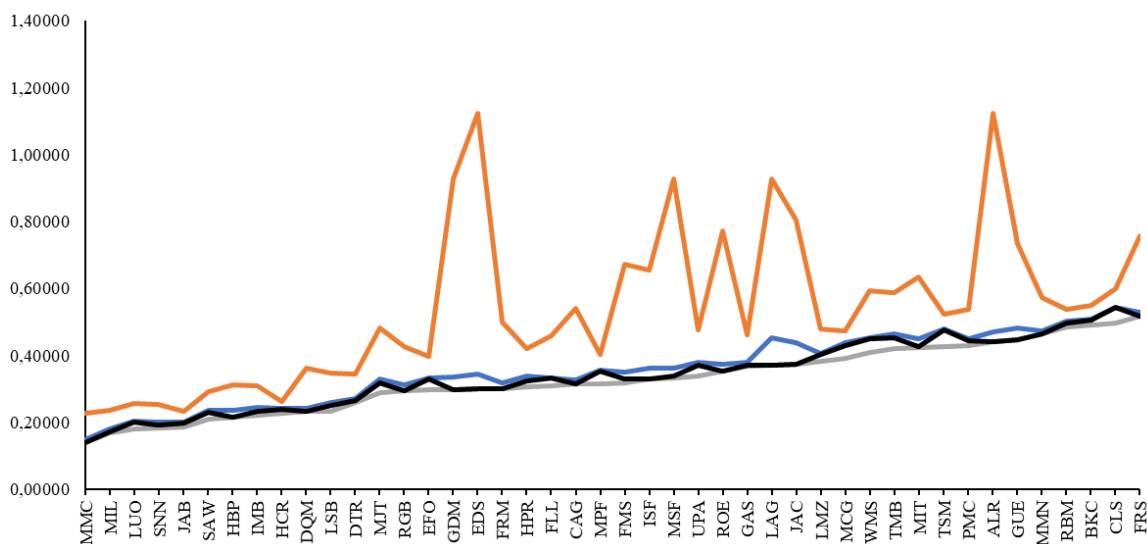


Figura 6A. Grau de redundância, classe S05 (50 simulações)

Legenda: cinza = menor; azul = média; preto: s00; laranja = maior.

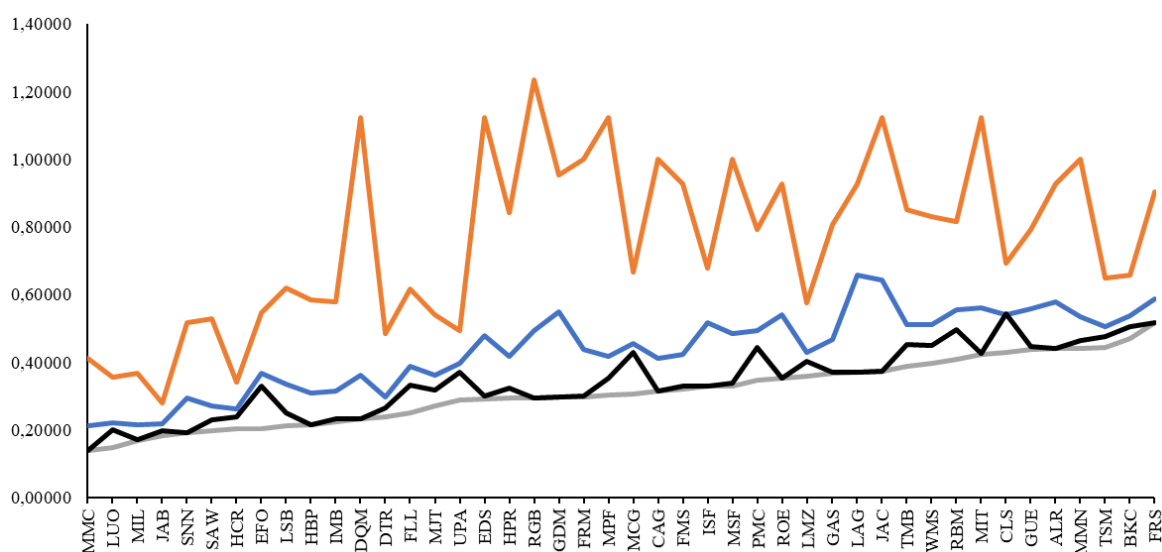


Figura 6B. Grau de redundância, classe S30 (50 simulações)

Legenda: cinza = menor; azul = média; preto: s00; laranja = maior.

As Figuras 6A e 6B são as que oferecem maior impacto visual demonstrativo da incerteza das medidas sociométricas relativas aos vértices. Por tais Figuras, mesmo com pequena redução do tamanho da amostra, a amplitude de variação é considerável e imprevisível (pelo menos para os analistas responsáveis pela investigação). Como alento, é razoável indicar que o valor real (de S00) do grau de redundância se encontra como regra entre a média e o menor valor (que aparentemente tendeu a ser a melhor estimativa do valor real).

Voltando-se agora para a distribuição de frequência e testando a sua aproximação com a Normal, foram computados o P-value para o teste Shapiro-Wicks, resumidos na Tabela 6, na qual a última coluna indica a quantidade de vértices (entre os 44 proeminentes), cuja distribuição para a grandeza indicada na linha sobreviveu ao teste de normalidade para os dois subconjuntos de 50 simulações (coluna **Classe**).

Tabela 6

Teste de Normalidade para as Características dos Vértices

Característica	Classe	P-value > 0,050
DGrau	S05	0
	S30	12
Centralidade por proximidade	S05	0
	S30	30
Centralidade por intermediação	S05	0
	S30	16
Grau de redundância	S05	1
	S30	12

O descasamento entre a média e a mediana inferidos nas Figuras 3A a 6B já prenunciaram os resultados negativos do teste de normalidade. Pela Tabela 6, isso é mais evidente nos testes para a classe S05, com uma única exceção – as distribuições de frequência das características estruturais dos vértices estão afastadas do padrão Gaussiano. À medida que se reduz o tamanho da amostra, os desvios se tornam

maiores, a média se aproxima da mediana e, em alguns casos, a distribuição toma um formato mais próximo da Normal (na Tabela 6, casos da classe S30). Contudo, a indefinição permanece: não foi encontrado um padrão que permitisse inferir qual vértice teria aquela ou outra característica mais próxima da distribuição de Gauss.

A única conclusão possível é que há um risco considerável e, pior, não quantificável pelos métodos intentados, de se utilizarem os dados de amostras para inferir as características de vértices na sua forma intervalar/razão. Resta verificar se há ao menos alguma estabilidade quando se trabalha com a escala ordinal.

Algumas indicações aparecem nas Figuras 3A a 6B, que ensejaram comentários prévios. Combinando-as, parece razoável afirmar que, apesar de algumas trocas de posição, elas se dão nas imediações da ordem estabelecidas em S00 (para verificar tal afirmativa, basta acompanhar o comportamento da sua linha nas figuras mencionadas, seguindo aproximadamente uma tendência).

Esse apontamento, indicativo de certo grau de estabilidade ordinal, sofre corroboração de um teste adicional, com resultados na Figura 7. Nela, são representados parcialmente (por economia) quatro Grupos de grau para a S00 (primeira linha) e para as 50 simulações da Classe S30 (depois do Label). A composição obedeceu à forma: Grupo A (em verde), sempre com os cinco vértices com maior grau em cada rede; Grupo B (em amarelo), da mesma forma, com os 10 vértices com maior grau, depois daqueles do Grupo A; Grupo C (em preto), os 29 intermediários, com grau entre aqueles pertencentes ao Grupo B e aqueles do grupo D; e, finalmente, o Grupo D (em branco), os demais 969 vértices, de menor grau. Apresenta-se, ainda, a computação da classificação na parte direita da Figura 7. Assim, JAB, que é A em S00, foi classificado corretamente quanto ao seu grau em 49 das 50 oportunidades (simulações) da Classe S30; por outro lado, EDS, último do Grupo C na Figura 7, foi classificado 14 vezes no grupo correto e 36 vezes no Grupo D.

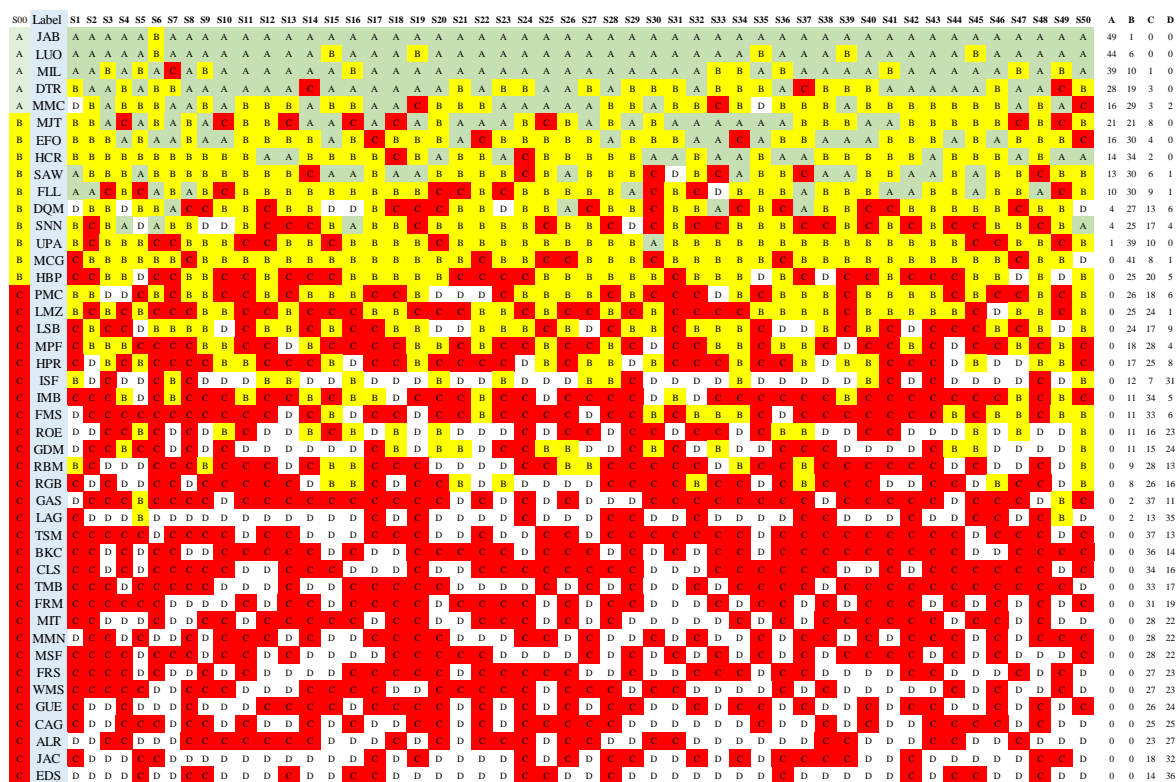


Figura 7. Agrupamentos Por Grau, S00 e S30 (Extrato Parcial)

A Tabela 7 é derivada da Figura 7, computando os agrupamentos.

Tabela 7

Agrupamento S00 & S30, Quanto ao Grau

		S30									
		Em termos absolutos					Em termos relativos				
		A	B	C	D	Soma	A	B	C	D	Soma
S00	A	176	65	7	2	250	0,70	0,26	0,03	0,01	1
	B	83	302	97	18	500	0,17	0,60	0,19	0,04	1
	C	0	187	736	527	1450	0,00	0,13	0,51	0,36	1
	D	0	0	798	47652	48450	0,00	0,00	0,02	0,98	1

Os dados indicam que, mesmo com uma redução de 30% (redução máxima testada), os agrupamentos são definidos corretamente em $176 + 303 + 736 + 47.652 = 48.866$ vezes entre as 50.650 oportunidades, o que indica a taxa de acerto próxima a expressivos 96%. A taxa é ainda maior (98%) para indicar os vértices **menos relevantes** do Grupo D. E são poucas as exceções de se **errar** por mais de um grupo (superior e inferior – **pular** mais de um grupo): isso aconteceu em nove oportunidades no Grupo A, 18 no Grupo B e em nenhuma oportunidade nos Grupos C e D. Daí, seria razoável afirmar que a identificação de vértices mais ou menos importantes na rede (através de seu grau), via uma amostra, demonstrou-se de risco **aceitável**, obviamente dependendo das alternativas e do grau de responsabilidade da classificação.

O poder classificatório da amostra também poderia ser testado nas outras características de estruturais. Todavia, o exercício foi considerado não recomendável, dado que os resultados satisfatórios, já obtidos pela medida mais simples, remetem à parcimônia: *pluralitas non est ponenda sine neccesitate* (a pluralidade sem necessidade deve ser evitada).

Conclusões e Recomendações

Finda uma pesquisa e avaliando-a, espera-se que o esforço despendido tenha sido compensador ao dar encaminhamento satisfatório para uma solução – sempre provisória, até melhores teorias, métodos, dados e análises – da questão proposta.

O primeiro ponto a merecer destaque é o favorecimento da pesquisa sociométrica censitária: mesmo não infensa de erros (Barbastefano et al., 2013, 2015; Wang, Shi, McFarland, & Leskovec, 2012), ela, por óbvio, não introduz (consideráveis) erros devido à amostragem, ainda mais e com frequência não probabilística. Somente quando a inviabilidade do censo for patente ou comprovada, deveria ser admitida a amostra. Mas, uma vez empregada amostra, é responsabilidade científica destacar as implicações do procedimento, e na medida do possível, procurar evidenciá-las – eventualmente, observando o método recorrido neste trabalho. Afinal, “entender a robustez das medidas básicas de rede é extremamente importante para assegurar a validade da análise de redes sociais” (Borgatti et al., 2006, p. 125).

No que diz respeito às características estruturais mais comuns das redes sócio-bibliométricas baseadas em coautoria, pesquisas que optam por amostra relativamente pequena e não aleatória, e depositam seus achados em variáveis como densidade, tamanho do componente maior e conectividade devem ser motivo de maior desconfiança do que aquelas com amostras aleatórias e maiores e que ficaram restritas a número de vértices, de ligações e de componentes. Já a utilização do grau médio, se necessário, apresentou-se menos arriscada do que o emprego de medidas alternativas para relacionar vértices e ligações.

Já em um plano mais prescritivo, aceitável para pesquisa de cunho tecnológico, parece razoável avaliar e apontar:

1. Uma fórmula para verificar a qualidade (precisão, exatidão) das características da rede foi testada e, até prova em contrário, parece razoável: faça simulações redutoras e estude as distribuições de frequência.
2. Se seus dados o permitirem, as simulações redutoras podem, em tese, ser utilizadas para inferir as grandezas mais simples da rede (número de vértices, de ligações e de componentes), desde que a amostra tenha um tamanho relativamente elevado (o que exige algum conhecimento da população; no caso, inferência sobre a quantidade total de itens publicados). O modelo linear para a regressão foi encontrado com boa performance para algumas das características da rede.
3. Se a pesquisa sócio-bibliométrica (baseada em coautoria) em questão pretende usar medidas de centralidade de vértices em modelação, e mesmo correlações, até as mais simples, como na apuração de capital social, empregue o censo. Em contrário, seus resultados deveriam ser postos em dúvida. Não se encontrou remédio para tal restrição, pelo menos nesta investigação e na literatura consultada.
4. Pelo menos, a identificação dos vértices mais relevantes, mediante a medida mais simples (seu número de parceiros, grau), em uma boa amostra parece guardar uma correspondência razoável com a população. Não parece a tratativa das mais atraentes, mas deve funcionar a contento.
5. Já há um conjunto considerável de pesquisas sócio-bibliométricas derivadas de coautorias. Algumas realmente perseguiram o censo e alcançam o campo escrutinado. Outras definem a **população** de forma tão estrita (aparentemente, para se credenciar como censo e/ou economizar esforços), que é evidente que ali se encontra uma amostra enviesada (mesmo que o viés seja **os periódicos de melhor qualificação Qualis** ou os **Programas de Pós-Graduação de nota Capes superiores**). Se o objetivo é fazer a caracterização estrutural do campo, tal estratégia deve ser recebida com reservas. Talvez, uma forma alternativa seria mapear o campo através do método de bola de neve via *curriculum* na Plataforma Lattes – proposta essa não explorada no trabalho.

Tendo avançado em alguns pontos, outros permanecem à disposição para trabalhos complementares. Por exemplo e por foco, não se testou a estabilidade de achados como homofilia, centro-periferia e **mundo pequeno**. As 300 simulações, já em Pajek ou em Ucinet, estão à disposição dos interessados na tarefa mediante solicitação.

Por fim, é imperativo falar nas limitações da pesquisa. Ela, e como outras, sempre esteve sujeita a erros operacionais, baldados os cuidados. Mas o ponto a ser ressaltado é de outra natureza, mais filosófica. Pela proposição de Popper (2014), um ou vários dados ou pesquisa não provam nada. Então, toda vez que aqui foi registrada a conveniência de uma característica ou de um método, isso, quando muito, é válido até que se tenham evidências do contrário. Talvez, não sobrevivam à próxima investigação.

Mas a relevância da pesquisa, o que a torna interessante e eventualmente merecedora de consideração pela comunidade acadêmica, está ainda em Popper (2014), na sua estratégia da falseabilidade: aqui, foram discorridas algumas refutações perturbadoras e, aí, sim, basta uma pesquisa para que ela possa almejar o *status* de contribuição.

Contribuições

- 1º autor: concepção; fundamentação teórica; metodologia; análise de resultados e redação.
2º autor: metodologia; desenvolvimento (simulação); análise de resultados e revisão.
3º autor: desenvolvimento (simulação); análise de resultados e revisão.
4º autor: desenvolvimento (simulação) e análise de resultados.

Referências

- Araújo, U. P., Mendes, M. de L., Gomes, P. A., Coelho, S. de C. P., Vinícius, W., & Brito, M. J. de. (2017). Trajetória e estado corrente da sociometria brasileira. *Redes - Revista Hispana para el Análisis de Redes Sociales*, 28(2), 97-128. <https://doi.org/10.5565/rev/redes.706>
- Barabási, A. L., Jeong, H., Nédá, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Application*, 311(3/4), 590-614. [https://doi.org/10.1016/S0378-4371\(02\)00736-7](https://doi.org/10.1016/S0378-4371(02)00736-7)
- Barbastefano, R. G., Souza, C., Costa, J. de S., & Teixeira, P. M. (2013). Impactos dos nomes nas propriedades de redes sociais: Um estudo em rede de coautoria sobre sustentabilidade. *Perspectivas em Ciências da Informação*, 18(3), 78-95. Recuperado de <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/download/1773/1194>
- Barbastefano, R. G., Souza, C., Costa, J. de S., & Teixeira, P. M. (2015). Influência da ambiguidade de nomes na centralidade de redes de coautoria. *TransInformação*, 27(3), 189-198. <http://doi.org/10.1590/0103-37862015000300001>
- Borgatti, S. P., Carley, K., & Krackhardt, D. (2006). Robustness of centrality measures under conditions of imperfect data. *Social Networks*, 28(2), 124-136. <http://doi.org/10.1016/j.socnet.2005.05.001>
- Borgatti, S. P., Everett, M. G., & Johnson, J. (2013) *Analysing social networks*. London: Sage.
- Burt, R. (1980). Actor interests in a social topology: Foundation for a structural theory of action. *Sociological Inquiry*, 50(2), 107-132. <https://doi.org/10.1111/j.1475-682X.1980.tb00380.x>
- Corley, E., Boardman, P. G., & Bozeman, B. (2006). Design and the management of multi-institutional research collaborations: Theoretical implications from two case studies. *Research Policy*, 35(7), 975-993. <https://doi.org/10.1016/j.respol.2006.05.003>
- Costenbader, E., & Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks*, 25(4), 283-307. [https://doi.org/10.1016/S0378-8733\(03\)00012-1](https://doi.org/10.1016/S0378-8733(03)00012-1)
- Granovetter, M. S. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, 91(3), 481-510. <https://doi.org/10.1086/228311>
- Hair, J. F., Jr., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise multivariada de dados* (6a ed.). Porto Alegre: Bookman.
- Hicks, D. (1998). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, 44(2), 193-215. <https://doi.org/10.1007/BF02457380>
- Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, 26(1), 1-18. [https://doi.org/10.1016/S0048-7333\(96\)00917-1](https://doi.org/10.1016/S0048-7333(96)00917-1)
- Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks*, 28(3), 247-268. <https://doi.org/10.1016/j.socnet.2005.07.002>
- Laumann, E. O., Marsden, P. V., & Prensky, D. (1983). The boundary specification problem in network analysis. In R. S. Burt & M. J. Minor (Eds.), *Applied network analysis* (pp. 18-34). London: Sage Publications.
- Lin, N. (1999). Building a network theory of social capital. *Connections*, 22(1), 28-51. Retrieved from <http://www.insna.org/PDF/Keynote/1999.pdf>

- Melin, G. (2000). Pragmatism and self-organization: Research collaboration on the individual level. *Research Policy*, 29(1), 31-40. [https://doi.org/10.1016/S0048-7333\(99\)00031-1](https://doi.org/10.1016/S0048-7333(99)00031-1)
- Milgram, S. (1967). The small world problem. *Psychology Today*, 1(1), 60-67. Retrieved from <http://snap.stanford.edu/class/cs224w-readings/milgram67smallworld.pdf>
- Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, 69(2), 213-238. <https://doi.org/10.1177/000312240406900204>
- Motta, G. da S. (2017). Editorial seção artigos tecnológicos: Como escrever um bom artigo tecnológico. *Revista de Administração Contemporânea*, 21(5). Retrieved from <http://www.scielo.br/pdf/rac/v21n5/1415-6555-rac-21-05-00004.pdf>. <http://dx.doi.org/10.1590/1982-7849rac2017170258>
- Newman, M. E. J. (2001). The structure of scientific collaboration networks. *PNAS*, 98(2), 404-409. <https://doi.org/10.1073/pnas.98.2.404>
- Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *PNAS*, 101(1), 5200-5205. <https://doi.org/10.1073/pnas.0307545100>
- Popper, K. (2014). *A lógica da pesquisa científica*. São Paulo: Editora Cultrix.
- Rigby, J., & Edler, J. (2005). Peering inside research networks: Some observations on the effect of the intensity of collaboration on the variability of research quality. *Research Policy*, 34(6), 784-794. <https://doi.org/10.1016/j.respol.2005.02.004>
- Robins, G., Pattison, P., & Woolcock (2004). Missing data in networks: Exponential random graph (p*) models for networks with non-respondents. *Social Networks*, 26(3), 257-283. <https://doi.org/10.1016/j.socnet.2004.05.001>
- Rossoni, L. (2014). Agência e redes mundos pequenos: Uma análise multinível da produtividade acadêmica. *Revista de Administração da Mackensie*, 15(1), 200-235. <http://doi.org/10.1590/S1678-69712014000100009>
- Rossoni, L., & Guarido, E. R., Filho (2007). Cooperação interinstitucional no campo da pesquisa em estratégia. *Revista de Administração de Empresas*, 47(4), 74-88. <http://doi.org/10.1590/S0034-75902007000400007>
- Sewell, W. H. (1992). A theory of structure: duality, agency, and transformation. *American Journal of Sociology*, 98(1), 1-29. <http://doi.org/10.1086/229967>
- Smith, J. A., & Moody, J. (2013). Structural effects of network sampling coverage I: Nodes missing at random. *Social Networks*, 35(4), 1-34. <http://doi.org/10.1016/j.socnet.2013.09.003>
- Stefano, D. de, Fuccella, V., Vitale, M. P., & Zaccarin, S. (2013). The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks*, 35(3), 370-381. <https://doi.org/10.1016/j.socnet.2013.04.004>
- Stefano, D. de, Giordano, G., Vitale, M. P. (2011). Issues in the analysis of co-authorship networks. *Quality & Quantity*, 45(5), 1091-1107. <https://doi.org/10.1007/s11135-011-9493-2>
- Stumpf, M. P. H., Wiuf, C., & May, R. M (2005). Subnets of scale-free networks are not scale-free: Sampling properties of networks. *PNAS*, 102(12), 4221-4224. <https://doi.org/10.1073/pnas.0501179102>
- Torman, V. B. L., Coster, R., & Riboldi, J. (2012). Normalidade de variáveis: Métodos de verificação e comparação de alguns testes não-paramétricos por simulação. *Revista do Hospital de Clínicas e*

da Faculdade de Medicina, 32(2), 227-234. Recuperado de <https://seer.ufrgs.br/hcpa/article/download/29874/19186>

Wang, D. J., Shi, X., McFarland, D. A., & Leskovec, J. (2012). Measurement error in network data: A re-classification. *Social Networks*, 34(4), 396-409. <https://doi.org/10.1016/j.socnet.2012.01.003>

Dados dos Autores

Uajara Pessoa Araujo
Av. Amazonas, 7675, 30441-008, Belo Horizonte, MG, Brasil
E-mail: uajara@yahoo.com.br

Fabício Molica de Mendonça
Av. Visconde do Rio Preto, S/N, Colônia do Bengo, 36307-352, São João del-Rei, MG, Brasil
E-mail: fabriciomolica@yahoo.com.br

Rita de Cássia Leal Campos
Av. Amazonas, 7675, 30441-008, Belo Horizonte, MG, Brasil
E-mail: rita.campos.adm@gmail.com

Lara Figueiredo e Silva
Av. Amazonas, 7675, 30441-008, Belo Horizonte, MG, Brasil
E-mail: larafigueiredo.s@hotmail.com