Estudo Experimental do Método Network Scale-Up em Diferentes Topologias de Grafos Aleatórios

João Pedro de Souza Olivo Tardivo[1], **Nicollas Mocelin Sdroievski**[1]

[1]Ciência da Computação
Universidade Estadual do Paraná (UNESPAR)
Apucarana – Paraná

1. INTRODUÇÃO

A estimação do tamanho de populações de difícil acesso, como grupos afetados por desastres naturais, portadores de doenças estigmatizadas ou membros de redes clandestinas, representa um desafio fundamental para a epidemiologia, as ciências sociais e a formulação de políticas públicas [1]. Métodos de enumeração direta são frequentemente inviáveis devido à natureza dispersa ou sigilosa desses grupos. Nesse contexto, a análise de redes sociais emerge como uma abordagem poderosa, permitindo inferir características de uma população inteira a partir das conexões locais de um subconjunto de seus membros [2]. Uma das técnicas mais proeminentes para essa finalidade é o Método Network Scale-Up (NSUM), uma abordagem de estimação indireta que utiliza Dados Relacionais Agregados (ARD) [3]. O método baseia-se em inquirir uma amostra de indivíduos sobre o tamanho de suas redes pessoais e quantos membros da população-alvo eles conhecem. A partir desses dados, dois estimadores são predominantemente utilizados para calcular a prevalência da população oculta: a Média das Razões (MoR), que pondera cada respondente igualmente, e a Razão das Somas (RoS), que implicitamente dá mais peso aos indivíduos mais conectados [4].

A precisão do NSUM, contudo, não é absoluta e sua performance é intrinsecamente dependente da topologia subjacente da rede social [5]. A escolha entre os estimadores MoR e RoS não é trivial, e trabalhos teóricos sugerem que a presença de heterogeneidade na distribuição de conectividade, como a existência de *hubs* e vértices altamente conectados, pode introduzir vieses e afetar a robustez de cada estimador de maneira distinta [6, 4]. Enquanto limitantes de erro analíticos fornecem uma base teórica, a validação empírica de seu comportamento em diferentes estruturas de rede é crucial para compreender sua aplicabilidade prática. O objetivo central deste trabalho é, portanto, conduzir um estudo experimental e comparativo para avaliar a acurácia e a robustez dos estimadores MoR e RoS sob diferentes topologias de grafos aleatórios. Por meio de simulação computacional, o desempenho dos estimadores é sistematicamente medido em três modelos de rede canônicos, cada um representando uma propriedade estrutural distinta: o modelo de Erdős-Rényi, como linha de base de uma rede homogênea e aleatória; o modelo de Barabási-Albert, para investigar o impacto da heterogeneidade de grau em redes *scale-free*; e o modelo de Watts-Strogatz, para analisar o efeito da alta clusterização local característica de redes de ”mundo pequeno”.

2. DESENVOLVIMENTO

As redes sociais podem ser modeladas formalmente como grafos [7, 8], onde os indivíduos são representados por vértices e as relações entre eles por arestas. O Método

Network Scale-Up (NSUM) emerge como uma abordagem indireta para estimar o tamanho de populações de difícil acesso [1, 9] em cenários onde a enumeração direta é inviável. O princípio fundamental reside na utilização de dados coletados por meio de perguntas diretas aos respondentes, assumindo que a rede pessoal de um respondente reflete a composição da população geral. A formulação básica [10] requer o número de pessoas conhecidas no grupo de interesse e o tamanho total da rede pessoal do respondente, ou seu grau. O método então calcula [11] a proporção de membros do grupo na população geral estimando essa proporção através da razão entre o número total de indivíduos do grupo conhecidos e o tamanho total das redes. Contudo, este estimador fundamenta-se em pressupostos como a mistura aleatória na rede [5], ausência de erros de recordação e acurácia na estimação do grau, sendo que desvios destes pressupostos podem introduzir vieses consideráveis [12].

Neste trabalho, analisa-se profundamente duas abordagens para calcular a proporção estimada da população oculta: o *Mean of Ratios* (MoR) e o *Ratio of Sums* (RoS) [4]. O estimador MoR representa uma abordagem no nível do indivíduo, calculando a proporção da população oculta dentro da rede pessoal de cada respondente e, em seguida, realizando a média aritmética simples dessas proporções sobre toda a amostra. A intuição é que cada respondente fornece uma estimativa independente, mas o método é sensível a respondentes com grau muito baixo [6]. Em contrapartida, o estimador RoS adota uma abordagem no nível da amostra, somando o número de contatos na população oculta reportado por todos os respondentes e dividindo pelo somatório dos graus. Isso resulta em um peso maior para os *hubs*, pois seus contatos contribuem mais para os totais [11], o que é vantajoso em redes heterogêneas.

Para avaliar a confiabilidade desses estimadores, utilizou-se a modelagem de grafos aleatórios [13, 14]. O modelo de Erdős–Rényi [15, 16] foi empregado como base, onde cada aresta é incluída com probabilidade independente, resultando em uma distribuição de grau binomial e pouca variabilidade entre os vértices. Dada a limitação desse modelo em representar redes reais heterogêneas [17], utilizou-se também o modelo de Barabási–Albert [18]. Este incorpora a ligação preferencial, onde novos vértices se conectam aos existentes com probabilidade proporcional ao grau atual, gerando redes *scale-free* com distribuição de lei de potência [19] e presença de vértices altamente conectados. Por fim, o modelo de Watts–Strogatz [20] foi incluído para capturar o fenômeno de ”mundo pequeno”, caracterizado por alto coeficiente de agrupamento e baixo comprimento médio do caminho, desafiando o arcabouço teórico que assume mistura aleatória perfeita e permitindo testar a robustez dos estimadores frente à alta coesão local [21].

A metodologia de pesquisa baseou-se em simulação computacional rigorosa. O fluxo de trabalho consistiu na configuração da rede sintética, definição da população oculta e sua prevalência real, amostragem de vértices simulando inquéritos, coleta de dados relacionais e cálculo das estimativas MoR e RoS. O desempenho foi quantificado por métricas como o erro relativo (normalizado para tratar super e subestimação simetricamente), a probabilidade de erro elevado (risco de exceder um limiar de 5%), o viés (tendência sistemática de erro) e a análise visual da distribuição via *boxplots*. Todos os experimentos foram implementados em Python, utilizando bibliotecas como NetworkX e NumPy, com sementes aleatórias fixas para reprodutibilidade e execução de 1.000 repetições por ponto de dados (50 grafos $\times$ 20 amostras).

Os resultados nas redes de Erdős-Rényi indicaram que, em uma topologia aleatória e homogênea, os estimadores MoR e RoS apresentam desempenho virtualmente idêntico. O erro médio decresceu consistentemente com o aumento da amostra para ambos, e as distribuições de erro foram indistinguíveis, com mediana próxima de 1.0 e redução simétrica na dispersão. A análise de risco mostrou que a probabilidade de erro acima de 5% é idêntica para os dois métodos, confirmando que em redes homogêneas não há vantagem prática na escolha de um sobre o outro.

Já nas redes de Barabási-Albert, a heterogeneidade estrutural expôs diferenças de desempenho. Contrariamente à expectativa teórica de um forte viés negativo para o MoR, os resultados de viés foram baixos para ambos. No entanto, a análise de risco demonstrou uma vantagem clara e consistente do estimador RoS. A probabilidade de o RoS produzir um erro significativo foi sistematicamente menor do que a do MoR, especialmente para amostras pequenas. Isso corrobora empiricamente os limitantes teóricos que sugerem que o RoS, ao ponderar pelos graus, lida melhor com a presença de *hubs*, tornando-se uma métrica mais confiável em redes *scale-free* [4].

Finalmente, nos experimentos com redes de Watts-Strogatz, buscou-se avaliar o impacto da clusterização. Mesmo variando a probabilidade de religação para alternar entre redes altamente ordenadas e mais aleatórias, ambos os estimadores se mostraram não viesados e com desempenho idêntico. As curvas de erro e probabilidade foram indistinguíveis até a quarta casa decimal. Isso permitiu concluir que a propriedade de ”mundo pequeno”e o alto agrupamento local não são os fatores que diferenciam a performance dos estimadores, reforçando que a divergência observada no modelo Barabási-Albert deve-se exclusivamente à heterogeneidade da distribuição de graus e não a correlações locais.

3. CONSIDERAÇÕES FINAIS

Este trabalho investigou a influência da topologia da rede no desempenho e na confiabilidade dos estimadores MoR e RoS do método NSUM. A análise empírica em três modelos de rede distintos permitiu isolar variáveis estruturais críticas. Os resultados obtidos nos modelos de Erdős-Rényi e Watts-Strogatz estabeleceram que, em topologias caracterizadas por uma distribuição de grau homogênea, o desempenho dos estimadores é virtualmente idêntico em todas as métricas de avaliação: erro médio, viés e risco de falha. Esta equivalência demonstrou que propriedades como a clusterização local não são o fator determinante na diferenciação dos estimadores.

A divergência de desempenho emergiu claramente apenas na topologia de Barabási-Albert. A análise neste modelo *scale-free* revelou que, embora a diferença no erro médio absoluto tenha sido modesta nos parâmetros testados, a análise de risco validou a superioridade teórica do estimador RoS. Neste cenário, o RoS apresentou uma probabilidade consistentemente menor de produzir estimativas com erro significativo, especialmente com amostras de menor tamanho. Conclui-se, portanto, que a superioridade de um estimador sobre o outro é dependente da estrutura da rede, sendo a heterogeneidade da distribuição de grau, especificamente a presença de *hubs*, o fator crucial que torna o estimador RoS tecnicamente mais robusto e confiável. Futuras investigações poderiam estender esta análise para modelos de blocos estocásticos e estratégias de amostragem mais complexas.

Referências

[1] T. H. McCormick and T. Zheng. Latent demographic profile estimation in hard-to-reach groups. *Ann. Appl. Stat.*, 6(4):1795–1813, December 2012.

[2] Matthew O. Jackson. *Social and Economic Networks*. Princeton University Press, 2010. ISBN 9780691148205.

[3] E. Breza, A. G. Chandrasekhar, T. H. McCormick, and P. Mengjie. Using aggregated relational data to feasibly identify network structure without network data. *American Economic Review*, 110(8):2454–84, August 2020. doi: 10.1257/aer.20170861. URL https://www.aeaweb.org/articles?id=10.1257/aer.20170861.

[4] Sergio Díaz-Aranda, Juan Marcos Ramirez, Mohit Daga, Jaya Prakash Champati, Jose Aguilar, Rosa Lillo, and Antonio Fernández Anta. Error bounds for the network scale-up method. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 498–509, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3736940. URL https://doi.org/10.1145/3711896.3736940.

[5] I. Laga, L. Bao, and X. Niu. Thirty years of the network scale-up method. *Journal of the American Statistical Association*, 116(535):1548–1559, 2021. doi: 10.1080/01621459.2021.1935267. URL https://doi.org/10.1080/01621459.2021.1935267. PMID: 37994314.

[6] J. P. Kunke, I. Laga, X. Niu, and T. H. McCormick. Comparing the robustness of simple network scale-up method (nsum) estimators, 2024. URL https://arxiv.org/abs/2303.07490.

[7] S. E. Fienberg. A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics*, 21(4):825–839, 2012. doi: 10.1080/10618600.2012.738106. URL https://doi.org/10.1080/10618600.2012.738106.

[8] Mark Newman. *Networks*. Oxford University Press, 2018. ISBN 9780198805090.

[9] R. Maltiel, A. E. Raftery, T. H. McCormick, and A. J. Baraff. Estimating population size using the network scale up method. *Ann. Appl. Stat.*, 9(3):1247–1277, September 2015.

[10] T. H. McCormick. *THE NETWORK SCALE-UP METHOD. The Oxford Handbook of Social Networks*. Oxford University Press, 2020.

[11] Sergio Díaz-Aranda, José Aguilar, Juan Marcos Ramírez, David Rabanedo, Antonio Fernández Anta, and .... Performance analysis of nsum estimators in social-network topologies. *The American Statistician*, 2025. doi: 10.1080/00031305.2024.2421361. URL https://www.tandfonline.com/doi/full/10.1080/00031305.2024.2421361.

[12] – authors not specified in snippet –. Estimating foreign born scientists and engineers (fbse). Technical report, Americas Data Hub, 2023. URL https://www.americasdatahub.org/wp-content/uploads/2025/04/Estimating_Foreign_Born_Scientists_and_Engineers_November_2023-final-report.pdf.

[13] S. Janson, T. Luczak, and A. Rucinski. *Random graphs*. John Wiley & Sons, 2011.

[14] László Lovász. *Large Networks and Graph Limits*, volume 60 of *Colloquium Publications*. American Mathematical Society, Providence, RI, 2012. ISBN 978-0-8218-9085-1. URL https://bookstore.ams.org/coll-60/.

[15] P. Erdös and A. Rényi. On random graphs i. *Publ. math. debrecen*, 6(290-297):18, 1959.

[16] Gerandy Brito, Ioana Dumitriu, and Kameron Decker Harris. Spectral gap in random bipartite biregular graphs and applications. *arXiv preprint*, arXiv:1804.07808, 2018. URL https://arxiv.org/abs/1804.07808.

[17] Mohammad Shahraeini. A comprehensive approach to synthetic distribution grid generation: Erdős–rényi to barabási–albert. *AUT Journal of Electrical Engineering*, 57(1):85–100, 2025. doi: 10.22060/eej.2024.23143.5591. URL https://eej.aut.ac.ir/article_5469.html.

[18] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[19] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[20] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998. doi: 10.1038/30918. URL https://www.nature.com/articles/30918.

[21] Mark D. Humphries and Kevin Gurney. Network 'small-world-ness': A quantitative method for determining canonical network equivalence. *PLOS ONE*, 3(4):e2051, 2008. doi: 10.1371/journal.pone.0002051. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0002051.