Explorando os Limites da Reprodutibilidade na Tarefa de Detecção de Comunidades em Modelos de Redes

Rainara Araújo Mateus
Department of Computer Science, Universidade Federal de Minas Gerais
Minas Gerais, Brazil
rainara.mateus@dcc.ufmg.br

Carlos H. G. Ferreira
Department of Computing and Systems, Universidade Federal de Ouro Preto, Minas Gerais, Brazil
chgferreira@ufop.edu.br

Ana Paula Couto da Silva
Department of Computer Science, Universidade Federal de Minas Gerais
Minas Gerais, Brazil
ana.coutosilva@dcc.ufmg.br

Abstract

Context: Network model-based studies are applied across various fields, including social media. *Problem*: Data availability and reproducibility are challenging due to restrictive data policies and the significant computational resources required to process large and complex networks. In this context, sampling techniques offer a viable alternative by selecting representative sub-networks that preserve the essential structural properties of the original network. Despite their potential, there is a dearth of studies investigating how sampling methods can generate networks at different scales and quantify their limitations in detecting communities, especially in conjunction with backbone extraction, a crucial step that can significantly affect the network's probabilistic properties. *Solution*: This paper evaluates the effectiveness of different sampling methods in improving the availability and reproducibility of network analysis, with a focus on community detection and backbone extraction. *SI Theory*: Our work is supported by Social Network Theory, which emphasizes relationships and connections among actors within a network over individual attributes. *Method*: In our research we quantify the limits, properties and scenarios in which smaller network versions can provide comparable communities structures to the original network. *Summarization of Results*: Our results show that certain sampling methods can effectively capture community structures even in reduced network representations. *Contributions and Impact in the area of SI:* This research facilitates the reproducibility and democratization of network studies and provides guidelines for creating networks at different scales that allow researchers to replicate certain studies.

Keywords

Reproducibility, Network community detection, Backbone extraction, Network sampling

1 Introdução

Modelos de rede têm um papel crucial no estudo em diversas áreas, como plataformas de mídias sociais, modelagem de discussões em fóruns, redes de coautoria acadêmicas, entre outras [16, 17, 57]. Nesse contexto, uma das tarefas mais importantes na análise de redes é a detecção de comunidades, que consiste na identificação de grupos de nós mais densamente conectados entre si do que com o restante da rede [18, 44]. Paralelamente, a extração de *backbones*, que envolve a filtragem das arestas menos relevantes, é frequentemente uma etapa essencial associada à detecção de comunidades [21, 32]. Esta técnica é particularmente importante porque permite uma visão mais clara e precisa dos fenômenos estudados. Isso é especialmente relevante em redes grandes e densas, em que ruídos podem obscurecer as estruturas de comunidades [9, 10, 33].

No entanto, a disponibilidade de dados com foco na reprodutibilidade de muitos estudos apresenta desafios significativos [24, 38]. Por exemplo, a análise de mídias sociais por meio de modelos de redes tem revelado diferentes padrões de interesse da sociedade sobre as interações de usuários e discussões [27, 37, 54]. Entretanto, plataformas como Twitter/X[1], Reddit[2] e LinkedIn[3] impuseram mais recentemente políticas de restrição que dificultam a coleta de dados [11, 31, 40]. Essas restrições afetaram drasticamente o acesso aos dados, tornando redes já modeladas e disponíveis na literatura, como as encontradas em repositórios como *Stanford Large Network Dataset Collection (SNAP)* e Network Repository[4], alternativas interessantes para estudos de redes.

Além disso, existe um desafio computacional não desprezível em processar tanto essas redes disponíveis quanto novas redes que eventualmente sejam disponibilizadas. Muitas dessas redes são tipicamente grandes e complexas, contendo até bilhões de nós e arestas, o que as torna difíceis de processar e analisar [2, 7, 53, 58]. Algoritmos de análise de redes, especialmente aqueles voltados para a detecção de comunidades e extração de *backbones*, exigem recursos computacionais substanciais [10, 43, 47, 58]. Infelizmente, nem todos os pesquisadores têm acesso à infraestrutura necessária para processar e analisar essas redes, recorrendo a ambientes colaborativos gratuitos, mas limitados, como Google Colab[5], Kaggle[6] e Microsoft Azure Notebooks[7]. Uma solução potencial para mitigar este problema é a disponibilização de versões menores e amostradas da rede original, permitindo que outros pesquisadores reproduzam estudos em diferentes escalas com recursos computacionais variados, mesmo que com alguma perda de fidelidade da estrutura e das comunidades inicialmente observadas. Diferentemente de trabalhos anteriores que aplicam amostragem, extração de *backbone* e detecção de comunidades separadamente, este trabalho avalia a combinação dessas técnicas para preservar as estruturas de comunidades e tornar a reprodutibilidade mais acessível. Essa abordagem é especialmente relevante para instituições com infraestrutura computacional limitada e busca fornecer diretrizes claras sobre quais métodos utilizar e em que contextos, promovendo a democratização e reprodutibilidade na análise de redes.

[1] https://developer.twitter.com/en/docs/twitter-api
[2] https://www.redditinc.com/policies/data-api-terms
[3] https://www.linkedin.com/legal/l/api-terms-of-use
[4] https://networkrepository.com/network-data.php
[5] https://colab.research.google.com/
[6] https://www.kaggle.com/
[7] https://notebooks.azure.com/

Conforme mencionado anteriormente, uma alternativa viável é o uso de técnicas de amostragem de redes, as quais são comumente associadas a várias tarefas, especialmente na detecção de comunidades [30, 39, 59]. Tais técnicas visam selecionar uma subrede representativa a partir de uma rede maior, preservando propriedades estruturais de interesse [23]. No entanto, faltam estudos que investiguem como métodos de amostragem podem ser explorados para gerar redes em múltiplas escalas, representando com fidelidade a rede original e focando na detecção de comunidades, com ou sem a extração de *backbones*. Quando a detecção de comunidades é associada à extração de *backbones* como uma etapa anterior, é necessário considerar que esses métodos são modelos probabilísticos que se baseiam em informações como grau, força e dados de vizinhança dos nós, os quais podem ser drasticamente afetados pelo processo de amostragem [21, 47]. Assim, é essencial compreender como diferentes métodos de amostragem, sejam eles mais elaborados e computacionalmente custosos (por exemplo, *Random Walk*) ou mais diretos e menos onerosos (por exemplo, *Random Node*), conseguem capturar subredes que permitam a reprodução eficaz dessa tarefa.

Nesse contexto, o objetivo deste estudo é explorar a eficácia de diferentes métodos de amostragem para melhorar a disponibilidade e, sobretudo, a reprodutibilidade de análises de redes em diferentes escalas, com foco na tarefa de detecção de comunidades, associada ou não à extração de *backbones*. Especificamente, buscamos entender os limites em que versões menores da rede podem ser utilizadas para obter resultados semelhantes às análises realizadas na rede originalmente modelada. Para isso, propomos uma metodologia baseada em estratégias de amostragem para quantificar os limites e cenários nos quais as redes podem ser reproduzidas em escalas menores, com alguma perda de informação em comparação com a rede original. Considerando que o volume massivo de arestas demanda bastante da memória principal [1], nossa abordagem visa explorar a viabilidade da detecção de comunidades tanto com quanto sem a extração de *backbones*, gerando amostras representativas.

Por meio de uma avaliação experimental em redes com diferentes características topológicas, observamos que alguns métodos de amostragem preservam bem as estruturas de comunidades, mesmo em redes substancialmente menores. Isso incentiva a reprodutibilidade de análises de redes e oferece diretrizes para criar versões em diferentes escalas, facilitando o acesso de pesquisadores com recursos computacionais limitados. Ao permitir a replicação de estudos em contextos com restrições computacionais, nosso trabalho contribui para a democratização da pesquisa em redes e promove uma ciência mais acessível. Os resultados indicam que versões reduzidas de redes podem manter informações estruturais essenciais, viabilizando análises escaláveis sem comprometer padrões fundamentais. Dessa forma, ampliamos o acesso a dados e metodologias, fortalecendo a disseminação da ciência e incentivando a colaboração entre pesquisadores com diferentes capacidades computacionais.

2 Fundamentos e Trabalhos Relacionados

2.1 Amostragem de Redes

Conjuntos de dados de grafos modernos são tipicamente grandes, trazendo desafios computacionais significativos [53, 58]. Com o crescimento no volume de dados [50], essas redes agregam diariamente mais interações entre entidades, tornando seu processamento custoso. Técnicas de amostragem são essenciais para lidar com esses grandes conjuntos de dados, permitindo a análise de subredes representativas de características topológicas de interesse [56].

Existem diversos métodos de amostragem na literatura [22]. Autores como Rozemberczki et al. [45] e S. Dizaji et al. [46] propuseram dividi-los em três grupos, conforme a estratégia de seleção: seleção de nós (*Node Selection*), seleção de arestas (*Edge Selection*) e amostragem baseada em travessia (*Traversal-based Sampling*). A seleção de nós envolve escolher um subconjunto de nós juntamente com todas as arestas que os conectam. Um exemplo dessa categoria é o *Random Node (RN)*, método que seleciona aleatoriamente nós da rede, dando a cada um deles igual probabilidade de ser escolhido, garantindo uma exploração não tendenciosa das diferentes partes da rede e proporciona uma seleção representativa [51]. Neste método, o parâmetro exigido é a fração de nós, que pode variar entre 0 e 1.

A seleção de arestas, por sua vez, escolhe um subconjunto de arestas e inclui os nós a elas incidentes. Já a amostragem baseada em travessia começa com um nó específico e expande a amostra seguindo um padrão determinado, como uma caminhada aleatória (em inglês, *Random Walk*), representando assim uma melhoria das estratégias anteriores ao incorporar elementos dinâmicos e adaptativos na seleção. O *Random Walk (RW)* é um método em que, a partir de um nó inicial, um vizinho é selecionado aleatoriamente, com a caminhada prosseguindo continuamente pela rede. Essa técnica é eficaz para alcançar regiões de difícil acesso em grandes redes e é especialmente útil em contextos de redes sociais [28, 60]. O parâmetro a ser definido para esse método é a fração de nós visitados durante a caminhada.

Outro método baseado em travessia é o *Forest Fire (FF)*, um processo iterativo que começa com um nó inicial e adiciona novos nós à amostra com base em uma probabilidade de conexão (p_f – a *probabilidade de queima*). Essa "queima" se espalha pela rede, conectando-se a outros nós com base em suas conexões, o que permite capturar a estrutura global da rede. Assim, áreas densamente conectadas têm maior probabilidade de serem amostradas, enquanto regiões menos densas tendem a ser menos exploradas [26]. Os parâmetros controlados nesse método incluem a fração de nós selecionados e a probabilidade p_f, ambos variam entre 0 e 1.

Vários estudos tentam entender o impacto dessas técnicas em diferentes aspectos das redes. Por exemplo, Stumpf et al. [51] revelaram que a seleção aleatória de nós em redes livres de escala não conserva bem a distribuição de graus. S. Dizaji et al. [46] avaliaram múltiplos métodos de amostragem, focando em métricas como grau e coeficiente de agrupamento. Aplicações de *Random Walks* em dados do *Facebook* destacam propriedades da rede e revelam abordagens enviesadas [20]. Zhao et al. [60] analisaram amostragens de redes sociais para definir uma abordagem complementar ao *Random Walk*. No contexto da detecção de comunidades, Leskovec and Faloutsos [26] exploraram técnicas de amostragem para aproximar as medidas da amostra às do gráfico original em redes sociais de larga escala, concluindo que *Random Walk* e *Forest Fire* tiveram melhor desempenho. Pons and Latapy [39] usaram *Random Walk* para quantificar a similaridade estrutural, enquanto Zhang et al.

[59] focaram no impacto de diversas técnicas na recuperação de *top-leaders* das comunidades.

Métodos como *Random Walk*, *Forest Fire* e *Random Node* têm mostrado potencial para preservar as estruturas das redes originais. No entanto, os trabalhos existentes em amostragem e extração de *backbones* abordam essas técnicas isoladamente, sem avaliar o impacto de sua combinação na detecção de comunidades. Essa integração é fundamental para identificar comunidades de forma mais precisa em redes grandes, densas e ruidosas [21]. A ausência de estudos que considerem essas abordagens em conjunto evidencia uma lacuna na literatura. Este trabalho busca preencher essa lacuna ao investigar como a amostragem pode ser explorada para gerar redes em múltiplas escalas que representem fielmente a estrutura original, com foco na detecção de comunidades, com ou sem a extração de *backbones*, visando aumentar a reprodutibilidade de estudos nessa área.

2.2 Detecção de Comunidades e Extração de *Backbones*

Detectar comunidades em redes é um problema amplamente estudado devido às suas diversas aplicações [4, 34, 44]. Existem várias definições de comunidades em redes na literatura, bem como diversos métodos de detecção de comunidades em redes, com abordagens variadas e amplamente discutidas na literatura [**?**]. Neste trabalho, focamos em comunidades estáticas e não sobrepostas, definidas como grupos de nós mais densamente conectados entre si do que com nós de outras comunidades [3]. Para esse propósito, adotamos o algoritmo de Louvain [7], que busca maximizar a *modularidade* — uma medida que avalia a qualidade das partições das comunidades, variando entre -0.5 e 1, com valores entre 0.3 e 0.7 indicando estruturas bem definidas [34]. Para uma revisão abrangente e detalhes sobre a implementação de outros métodos, recomendamos **?**].

Além da modularidade, diversas métricas podem ser utilizadas para avaliar a qualidade das comunidades detectadas [44]. Entre elas, a Informação Mútua Normalizada (*Normalized Mutual Information* - NMI) é amplamente empregada na literatura para comparar partições de comunidades e avaliar sua consistência [14, 44]. A NMI é uma medida não linear baseada na entropia de Shannon da teoria da informação e estima o quanto uma partição de comunidades pode informar sobre a outra. Seu valor varia de 0, indicando que todas as comunidades mudaram completamente, até 1, quando as comunidades são idênticas [48]. O cálculo da NMI entre duas partições X e Y é dado por:

$$NMI(X,Y) = \frac{\sum_x \sum_y P(x,y) \log \frac{P(x,y)}{P(x)P(y)}}{\sqrt{H(X)H(Y)}}, \quad (1)$$

onde $P(x)$ é a probabilidade de um nó ser atribuído à comunidade x, $P(x,y)$ é a probabilidade de um nó pertencer simultaneamente às comunidades x em X e y em Y, e $H(X) = -\sum_x P(x) \log P(x)$ representa a entropia de Shannon para a partição X. A NMI captura não apenas a coincidência direta entre as comunidades, mas também as relações informacionais entre as partições.

Mais recentemente, vários estudos têm destacado como a complexidade das interações entre usuários em grandes redes pode obscurecer as verdadeiras estruturas de comunidades, sendo necessário o uso de técnicas de extração de *backbones*. A extração de *backbones* visa remover arestas irrelevantes, mantendo apenas aquelas salientes, ou seja, importantes para o fenômeno em estudo [21, 32]. Em termos simples, o objetivo central das técnicas de *backbone* é obter uma versão reduzida, porém mais representativa daquilo que se deseja estudar na rede [10]. No contexto da detecção de comunidades, o uso da extração de *backbones* tem se mostrado bastante benéfico, preservando apenas as arestas relevantes para um dado estudo e melhorando a descoberta de comunidades mais estruturadas e representativas do problema alvo [17, 55].

A combinação de extração de *backbones* e detecção de comunidades tem sido aplicada em diversos contextos. Em plataformas de mídias sociais, esses métodos revelam padrões de disseminação de informação em redes de *retweets* e co-*retweets* no Twitter/X [12, 27, 37], redes de compartilhamento no WhatsApp e Telegram [35, 54], redes de co-*commenters* no Instagram [17], e comunidades de usuários em *subreddits* [36]. Essas técnicas também têm sido utilizadas em outros domínios, como a Darknet [49], redes de coautoria científica [25], redes fonológicas [55], sistemas de *tags* em plataformas de perguntas e respostas [57], e redes de transações financeiras [42].

Para isso, esses estudos utilizam métodos de *backbone* como o *Disparity Filter*, *Polya Urn Filter*, *Tripartite Backbone Extraction*, entre outros, que são tipicamente modelos probabilísticos que utilizam um modelo nulo para estimar o peso esperado de uma aresta. Esse modelo nulo é uma rede teórica construída com base nas propriedades estruturais dos nós e das redes, como o grau dos nós ou a distribuição das arestas, sem considerar as conexões específicas da rede original. A partir dessa construção teórica, é possível identificar quais arestas na rede original possuem pesos significativamente maiores do que o esperado pelo modelo nulo, indicando que são arestas relevantes para a estrutura da rede. Esses métodos são amplamente utilizados e têm se mostrado eficazes em fornecer estruturas de comunidades mais claras e mais relacionadas a um dado problema em estudo [17, 21, 27, 32, 47].

No entanto, pouco se sabe sobre o comportamento desses métodos quando aplicados a amostras de uma rede maior. Entender tal aspecto permitiria a criação de redes reduzidas e amostradas, facilitando estudos de detecção de comunidades em diferentes escalas com alguma fidelidade à rede original, o qual é o foco deste estudo. Tendo em vista o contexto apresentado anteriormente, nosso trabalho prevê a investigação de métodos de amostragem aplicados em redes para avaliar a efetividade do uso de amostras na representação de redes completas e *backbones*. Assim, nossa abordagem propõe entender a viabilidade da disponibilização de amostras que requerem menos recursos computacionais para a geração de comunidades com estrutura semelhante à rede original.

3 Metodologia

A Figura 1 apresenta a metodologia proposta. Primeiramente, envolve a definição de um cenário de interesse, tomando a rede originalmente modelada e selecionando um conjunto potencial de métodos de amostragem de redes para entender os impactos na tarefa de detecção de comunidades e encontrar o mais adequado para a rede em questão. Esta rede pode ser derivada de estudos anteriores ou pode ser uma rede de um novo estudo em que há interesse em compartilhá-la para estudos futuros. Em ambos os casos, o interesse reside em compreender os limites de representatividade

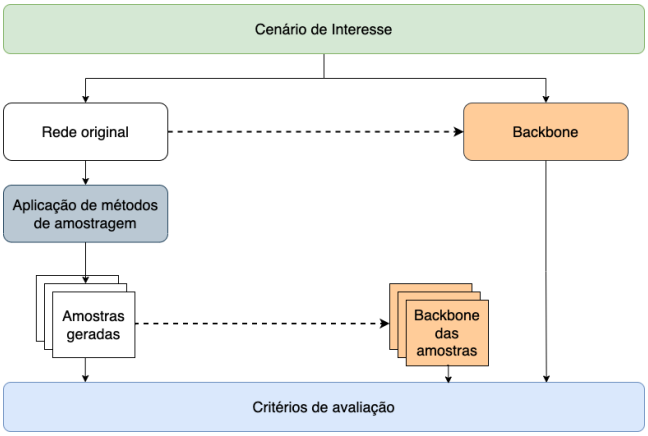

Figura 1: Fluxo da metodologia proposta.

das sub-redes amostradas na tarefa alvo, focando em aumentar a reprodutibilidade. Além disso, também consideramos o uso ou não da extração de *backbone* como uma etapa preliminar de detecção de comunidades. Isso é feito para entender o quanto métodos para esse propósito sofrem com o processo de amostragem, já que muitos deles são modelos probabilísticos que operam sobre a distribuição dos graus, pesos das arestas e vizinhança na rede, estando, portanto, sujeitos a impactos devido à amostragem.

Em seguida, diferentes níveis de amostragem são gerados usando cada método de amostragem de rede, variando o tamanho da amostra da rede, desde redes pequenas até aquelas mais próximas do original, conforme os parâmetros do método. Ressalta-se, no entanto, que os métodos de amostragem devem ser relacionados à tarefa alvo. Como focamos na detecção de comunidades, métodos de amostragem apropriados para esse propósito devem ser investigados na literatura e empregados. Como mencionado na Seção 2, métodos aplicáveis a nossa tarefa incluem a amostragem aleatória de nós (em inglês, *Random Node*) (RN), útil para tarefas gerais onde cada nó tem igual probabilidade de ser incluído na amostra, fornecendo uma representação imparcial da rede [51]; *Forest Fire* (FF), eficaz para tarefas como a detecção de comunidades, pois captura a estrutura global e regiões locais densamente conectadas [26]; e a amostragem por caminhada aleatória (*Random Walk*) (RW) adequada para tarefas que requerem a exploração de componentes e regiões conectadas da rede [60]. Em todos os casos, os parâmetros dos métodos, discutidos na Seção 2, foram ajustados de 0 a 1, em incrementos de 0.1.

Por fim, o objetivo final é realizar uma avaliação de desempenho da perda de informação, tomando como entrada a rede original, com ou sem extração de *backbone* como etapa preliminar da detecção de comunidades, bem como as sub-redes amostradas e seus *backbones*, para entender os limites da amostragem na tarefa final conforme a rede e os métodos usados. Como mencionado, a extração de *backbone* tem sido favorável para a detecção de comunidades, mas impõe custos computacionais, bem como riscos de não capturar as arestas adequadas, por serem modelos probabilísticos que podem sofrer com a estrutura de rede incompleta. Já os critérios de avaliação incluem métricas que dependem da tarefa alvo. Nosso foco aqui está na detecção de comunidades, onde métricas como Informação Mútua Normalizada (NMI) podem ser usadas para quantificar a similaridade das comunidades nas redes originais e aquelas amostradas [44]. No entanto, se a tarefa mudar, como medir a influência dos nós, métricas como centralidade de grau, centralidade de intermediação e proximidade podem ser usadas para observar a correlação da influência dos nós entre as redes original e amostrada, com e sem extração de *backbone* [46]. Para tarefas de predição de arestas, precisão, revocação e F1-score podem ser empregados para avaliar a precisão das arestas previstas nas redes amostradas em comparação com a rede original [8, 44]. Além disso, testes estatísticos podem ser utilizados para identificar diferenças significativas entre os métodos de amostragem. Dessa forma, esperamos estimar os efeitos e limites do uso de técnicas de amostragem, fornecendo informações sobre como diferentes estratégias de amostragem e métodos de extração de *backbone* impactam a análise de tarefas complexas em redes.

4 Avaliação Experimental

Nesta seção, apresentamos uma avaliação experimental, descrevendo as redes modeladas, os métodos de *backbone* empregados e o uso do algoritmo *Louvain* para detecção de comunidades [7]. Também detalhamos os métodos de amostragem e os critérios de avaliação adotados.

4.1 Modelagem das Redes e dos *Backbones*

- **WhatsApp:** A rede é modelada como um grafo não direcionado e ponderado $G = (V, E)$, onde V representa os usuários que postaram uma mensagem em um grupo na plataforma, e E representa os usuários que postaram mensagens nos mesmos grupos. Uma aresta conecta dois nós (usuários) se eles postaram pelo menos uma mensagem em comum em diferentes grupos observados, com o peso w_{ij} representando o número de mensagens que postaram em comum. O foco do trabalho foi estudar a disseminação de informações a partir do trabalho realizado em [35], cujo conjunto de dados foi gentilmente cedido pelos autores. Eles empregaram o método *Disparity Filter* [47] para extração de *backbone*, visando capturar comunidades de usuários que disseminam informações entre diferentes grupos.
- **Instagram:** Esta rede consiste em dados de comentários em postagens de perfis políticos no *Instagram* durante a semana das eleições em 2018 [17, 21], cujo conjunto de dados foi gentilmente cedido pelos autores. A rede é representada como um grafo não direcionado e ponderado $G = (V, E)$, onde V são os usuários que comentam nas postagens e E representam conexões entre usuários que comentam na mesma postagem, com o peso w_{ij} correspondendo ao número de postagens que receberam comentários de ambos os usuários. Um dos métodos explorados pelos autores foi o *Gloss Filter* [41], que se mostrou eficaz para a extração de *backbone* nesta rede, para revelar e caracterizar as comunidades de co-comentadores engajados em discussões políticas.
- **Tags Stack Overflow:** Esta rede representa as co-ocorrências de tags em perguntas nos fóruns online do *Stack Overflow* [5, 6, 19], cujo conjunto de dados foi disponibilizado pelos autores de [5, 6]. O grafo $G = (V, E)$ é representado por V como as tags e E como as arestas que conectam duas tags se elas co-ocorreram em pelo menos uma pergunta, com o peso w_{ij} representando o número de perguntas nas quais co-ocorreram. O método *Disparity Filter* [47] é proposto para ser usado na extração de *backbone* [15, 29], visando revelar grupos de tags representativos de comunidades que persistentemente co-ocorrem em questões.

4.2 Seleção dos Métodos de Amostragem

Diversos métodos para amostragem de redes são discutidos na literatura [23, 26]. Focamos na detecção de comunidades, tarefa amplamente relevante na análise de redes sociais [4], selecionando métodos reconhecidos por sua eficácia nesse contexto. Em redes reais, a distribuição de graus é altamente heterogênea, e amostragens tendenciosas a nós de alto grau podem distorcer a estrutura e ocultar comunidades menores [23, 46]. Métodos que evitam esse viés são essenciais para obter amostras representativas e análises precisas [52]. Em razão disso, escolhemos *Random Node*, *Forest Fire* e *Random Walk* por serem amplamente utilizados e reconhecidos na literatura por equilibrar representatividade estrutural e eficiência computacional em detecção de comunidades, como explicado na Seção 2.

4.3 Critérios de Avaliação

Para avaliar a extensão com que uma estrutura de comunidade é mantida em uma rede e em uma amostra, usamos a Informação Mútua Normalizada (NMI) explicada na Seção 2. Neste caso, somente os nós em comum nas duas redes (original e amostra) são considerados.

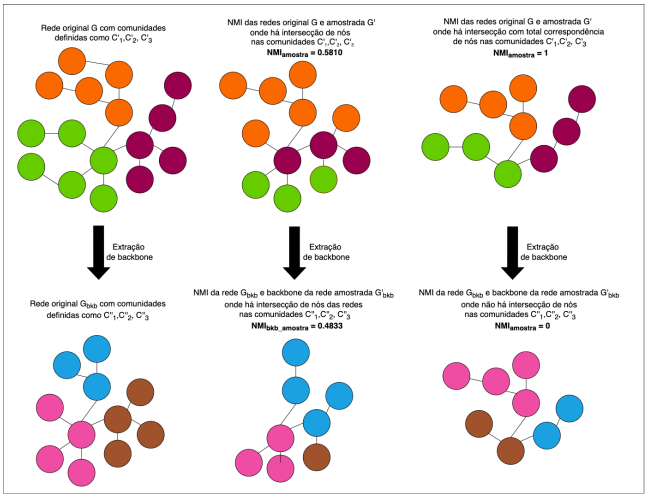

Figura 2: Exemplificação da avaliação da detecção de comunidades de uma rede, suas amostras e seus respectivos *backbones*.

A Figura 2 ilustra a avaliação de uma rede, onde apresentamos por cores os grupos de comunidades da rede original, amostras, *backbone* da rede original e *backbone* das amostras. Inicialmente, é realizada a detecção das comunidades da rede original através do algoritmo Louvain [7]. Em seguida, são apresentadas amostras do exemplo da rede original, cujo NMI é calculado a partir das intersecções de nós entre a rede original e amostradas. No primeiro caso, temos uma correspondência parcial de nós entre as comunidades da rede original e amostrada, ocasionando na obtenção de NMI cujo valor é 0.5810. Já na segunda amostra da rede original, observa-se total correspondência dos nós da intersecção das redes com as comunidades obtidas e, como consequência, acarretando no valor de NMI igual a 1.

Para a estrutura dos *backbones* apresentadas pelas setas, utiliza-se os nós de intersecção entre o *backbone* da rede original e o *backbone* da amostra. A primeira estrutura de *backbone* da amostra apresenta correspondência parcial de nós entre as comunidades do *backbone* da rede original e do *backbone* da amostra, gerando o NMI de valor 0.4833. Para o segundo cenário, a estrutura de *backbone* da amostra não apresenta nenhuma correspondência de nós da intersecção das redes entre as comunidades do *backbone* da rede original e do *backbone* da amostra, gerando valor de NMI igual a 0.

Essa avaliação é importante, porque se a NMI for igual a 0, isso indica que até mesmo a rede original tem limitações em reproduzir consistentemente as comunidades encontradas, já que o algoritmo de Louvain é uma heurística [7]. Portanto, tal aspecto deve ser considerado no processo de avaliação para evitar atribuir tais limitações aos métodos de amostragem. Em seguida, pegamos a rede original e aplicamos cada um dos métodos de amostragem propostos, variando seus respectivos parâmetros. Como métodos de amostragem sofrem de não-determinismo, para cada método e cada parâmetro, geramos dez amostras variando a semente e aplicamos o algoritmo Louvain a cada configuração. Então, calculamos novamente a média e o desvio padrão da NMI entre a estrutura de comunidade de cada configuração (método e parâmetro) e uma das estruturas de comunidade da rede original. É desejável que esse NMI médio seja similar ao da rede original, mostrando que a mesma estrutura de comunidades ou próxima é recuperada.

No caso da rede com *backbone*, iniciamos o procedimento extraindo o *backbone* da rede original. Em seguida, extraímos o *backbone* das amostras da rede original. Em seguida, usamos o mesmo método de extração de *backbone* dos cenários definidos na seção anterior. Ou seja, aplicamos a extração de *backbone* a cada uma das dez amostras de cada configuração. Novamente, executamos o algoritmo Louvain e calculamos a média e o desvio padrão da NMI para as estruturas de comunidade encontradas nos *backbones* extraídos das redes amostradas sobre o *backbone* da rede original. Mais uma vez, comparamos os valores do NMI médio entre o *backbone* na rede original e os *backbones* das configurações amostradas para entender a perda de reprodutibilidade devido à amostragem, agora considerando a etapa de extração de *backbone*.

Com todos esses dados, é possível avaliar como métodos de amostragem e parametrizações diferentes, com ou sem o uso de *backbones*, capturam a estrutura de comunidade originalmente observada em diferentes escalas, permitindo entender os limites da reprodutibilidade de sub-redes e identificar o melhor método e parametrização. A avaliação experimental está disponível em um repositório público [8].

5 Resultados

Essa seção apresenta nossos resultados e discute os principais achados. Nas tabelas com as características topológicas das redes, $\#V$ representa o número de nós em cada rede, $\#E$ é o número de arestas, d é a densidade das redes, $\hat{k}$ é o grau médio dos nós, CMA é o coeficiente médio de agrupamento dos nós, $\#\,Com.$ é o número médio de comunidades encontradas e Q é a modularidade da partição conforme o algoritmo de Louvain [7, 34]. Para os valores não

[8] https://github.com/rainara-araujo/network-sampling-communities

determinísticos, como o número de comunidades, modularidade e NMI, a média e o desvio padrão são apresentados.

5.1 WhatsApp

A Tabela 1 apresenta a caracterização topológica e das comunidades da rede WhatsApp e do seu *backbone*. Os números mostram que o *backbone* reduz a complexidade da rede, diminuindo o número de nós, arestas, e, sobretudo, melhorando a modularidade média da rede e aumentando o NMI médio das várias execuções do algoritmo, conforme explicado na Seção 3.

Tabela 1: Caracterização Topológica e de Comunidades da Rede e do *Backbone* do WhatsApp.

Rede	# V	# E	$\hat{k}$	CMA	d	# Com.	Q	NMI
Original	4341	221002	101.82	.61	.023	9.4±.84	.32±.004	.45±.03
Backbone	1154	15993	27.71	.54	.024	7.8±.8	.45±.004	.59±.03

O NMI em ambos os casos revela que, mesmo entre diferentes execuções do algoritmo Louvain, ainda existe uma variação nas comunidades devido à topologia natural da rede. É importante destacar a relevância deste NMI médio, considerando a instabilidade do algoritmo Louvain, que é uma meta-heurística [7]. A variação das soluções produzidas pelo Louvain já afeta a reprodutibilidade dos resultados, uma vez que diferentes pesquisadores podem obter divisões de comunidades distintas, mesmo utilizando o mesmo algoritmo e os mesmos dados. Isso pode ser problemático para a validação e comparação de estudos. Portanto, qualquer comparação deve ser feita em relação ao NMI médio de várias execuções das redes completas que se pretende amostrar. Por esse motivo, apresentamos a média e o desvio padrão para o número de comunidades, a modularidade e o NMI, visando uma análise mais justa dos resultados subsequentes.

Focando nos resultados do método *Random Node (RN)* da Tabela 2, nota-se que o NMI varia bastante para amostras pequenas até amostras grandes representadas pela fração de nós (*Frac. Nós*), especialmente para o caso em que o *backbone* não é usado. Para a rede original, por exemplo, amostras compostas por 60% dos nós na rede já fornecem uma aproximação interessante, conseguindo recuperar mais de dois terços (.33/.45) do NMI original, com redução de aproximadamente de 64% das arestas da rede original, contribuindo para a redução do custo computacional dessa rede. Para o *backbone*, a flexibilidade em usar amostras menores é muito maior. Nota-se que, acima de 60% dos nós, a perda do NMI médio (.50) em relação ao original (.59) é bem menor, e nesse caso houve redução de cerca de 64% das arestas do *backbone* da rede original. Deve-se também considerar que o *backbone* representa uma estrutura bem menor do que a da rede original em termos do número de nós e arestas. Portanto, considerando esse método, a extração de *backbone* provê subamostras bem mais interessantes, com capacidade de se aproximar significativamente da rede original.

Tabela 2: NMI nas amostras usando o *RN* no WhatsApp.

Frac. Nós	.1	.2	.3	.4	.5	.6	.7	.8	.9
Original	.24±.03	.23±.03	.25±.03	.27±.03	.28±.04	.33±.03	.35±.03	.39±.05	.40±.03
Backbone	.38±.22	.45±.06	.46±.05	.44±.04	.47±.04	.50±.05	.51±.05	.54±.04	.57±.02

Tabela 3: NMI nas amostras usando o *FF* no WhatsApp.

p_f / Frac. Nós	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	.48±.05	.48±.05	.48±.05	.46±.08	.48±.06	.44±.05	.48±.06	.46±.05	.46±.06
0.2	.52±.04	.53±.05	.53±.05	.53±.08	.53±.04	.54±.03	.54±.07	.53±.04	.54±.05
0.3	.55±.03	.58±.02	.56±.05	.55±.03	.56±.03	.60±.03	.58±.04	.55±.06	.58±.03
0.4	.55±.03	.55±.03	.56±.03	.56±.05	.57±.03	.55±.05	.57±.03	.55±.03	.56±.03
0.5	.53±.05	.52±.04	.53±.04	.52±.05	.53±.04	.53±.02	.52±.06	.55±.03	.53±.03
0.6	.50±.03	.47±.05	.50±.04	.48±.05	.49±.05	.48±.04	.49±.03	.50±.03	.52±.02
0.7	.45±.03	.44±.03	.46±.03	.45±.03	.47±.05	.44±.05	.47±.04	.47±.04	.50±.04
0.8	.44±.03	.46±.04	.44±.03	.43±.04	.45±.03	.45±.04	.45±.04	.47±.05	.47±.04
0.9	.41±.04	.43±.04	.44±.03	.45±.04	.47±.03	.43±.04	.47±.02	.44±.03	.46±.03

Na Tabela 3, temos os valores referentes de NMI obtidos pelo método *Forest Fire (FF)*, baseado na exploração de nós proporcional à *probabilidade de queima* (p_f). A linha da tabela representa p_f , enquanto a coluna representa a fração de nós amostrados (*Frac. Nós*). No caso do *WhatsApp*, esse método gerou amostras com substancial similaridade com os nós das comunidades originais, mesmo para amostras de 10% dos nós, que tiveram uma redução de cerca de 90% das arestas da rede original, sendo satisfatórias para fornecer uma representação semelhante das comunidades da rede original, considerando o NMI médio.

No caso do *backbone* das amostras, os resultados também foram similares (Tabela 4). No entanto, neste caso, a seleção do nó é afetada por maiores *probabilidades* p_f para gerar comunidades semelhantes ao *backbone* original. Nota-se que amostras com 10% dos nós, independente da probabilidades p_f, apresentam valores de NMI próximos ao NMI original do *backbone*. Com amostras de 10% obteve-se uma redução de até 85% das arestas, evidenciando uma representatividade próxima com poucos nós.

Tabela 4: NMI nos backbones das amostras usando o *FF* no WhatsApp.

p_f / Frac. Nós	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	.47±.06	.46±.05	.45±.05	.45±.05	.47±.06	.43±.07	.47±.05	.45±.04	.44±.06
0.2	.48±.05	.51±.06	.50±.04	.50±.06	.48±.04	.51±.05	.49±.04	.50±.04	.50±.03
0.3	.51±.05	.53±.05	.53±.04	.54±.04	.54±.05	.53±.03	.55±.02	.53±.05	.53±.04
0.4	.55±.03	.56±.05	.56±.05	.54±.02	.55±.03	.56±.02	.56±.04	.55±.03	.55±.02
0.5	.58±.02	.57±.03	.56±.02	.57±.04	.55±.04	.58±.03	.57±.03	.57±.02	.59±.07
0.6	.57±.02	.58±.03	.56±.02	.58±.03	.56±.04	.57±.03	.58±.02	.58±.04	.57±.04
0.7	.58±.02	.57±.03	.59±.03	.58±.03	.59±.02	.57±.02	.59±.03	.58±.03	.57±.04
0.8	.59±.01	.59±.02	.58±.02	.58±.03	.60±.02	.57±.03	.59±.02	.61±.03	.60±.02
0.9	.59±.02	.60±.02	.59±.04	.60±.03	.61±.03	.59±.03	.60±.04	.60±.03	.60±.03

Nos resultados do *Random Walk (RW)* no *WhatsApp* apresentados na Tabela 5, pela análise do NMI nas amostras da rede original, este método forneceu estruturas com membros de comunidade similares às suas comunidades da rede originais. Nas amostras geradas com tamanho a partir de 10%, as quais apresentaram redução de cerca de 90% das arestas da rede original, temos NMI que supera a média do NMI original. Sobre o *backbone* das amostras, o *WhatsApp* também apresentou resultados significativos onde amostras com tamanho de 50% podem representar todas as comunidades do *backbone* da rede original com redução de arestas de 6%. Entretanto, para amostras de tamanho de 30% dos nós a redução de arestas é cerca de 31% das arestas, revelando ser capaz de fornecer versões da rede em menores escalas representativas e menos custosas computacionalmente.

Em suma, para o *WhatsApp*, a adoção de métodos de amostragem baseados em exploração de nós, *FF* e *RW*, demonstra que é possível obter amostras que preservam a estrutura das comunidades originais de maneira eficaz. A utilização da extração de *backbone*

Tabela 5: NMI nas amostras usando o *RW* no WhatsApp.

Frac. Nós	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	.9
Original	.49±.02	.51±.04	.57±.03	.58±.02	.52±.05	.50±.03	.48±.04	.45±.02	.45±.03
Backbone	.46±.04	.50±.04	.52±.04	.55±.03	.59±.03	.57±.02	.59±.02	.57±.03	.61±.03

Tabela 6: Caracterização Topológica e de Comunidades da Rede e do *Backbone* do Instagram.

Rede	# V	# E	$\hat{k}$	CMA	d	# Com.	Q	NMI
Original	27615	125415544	9083.15	.73	.33	4.4±.52	.18±.003	.82±.06
Backbone	17498	256276	29.29	.27	.0017	5.1±.32	.28±.0027	.89±.02

após a amostragem melhora ainda mais a modularidade e a similaridade das comunidades detectadas em relação à rede original. Independentemente da complexidade do método de amostragem, os resultados mostram que é possível capturar a essência das comunidades da rede original, seja ela com *backbone* ou não, utilizando amostras menores e em diferentes escalas. Considerando o percentual de remoção de arestas da rede original, pode-se observar que mesmo com amostras relativamente pequenas e independente das probabilidades p_f, *FF* consegue representar as comunidades da rede original de forma fiel.

5.2 Instagram

Os resultados apresentados na Tabela 6 destacam a complexidade topológica da rede original do Instagram, que conta com 27.615 nós e 125.415.544 arestas. Isso ressalta a importância de fornecer versões representativas das redes, que, embora simplificadas, permitem, de preferência, a reprodutibilidade de estudos com algum grau de perda de informação. A complexidade da rede original impõe um custo computacional significativo. Mesmo após a redução com a extração de *backbone*, a rede permanece considerável, com 17.498 nós e 256.276 arestas.

A extração de *backbone* melhora a estrutura de comunidades e o NMI médio, fornecendo uma visão aprimorada da estrutura das comunidades. Especificamente, a rede *backbone* resulta em um aumento do NMI médio de 0.82 para 0.89, indicando uma representação mais fiel das comunidades originais. Além disso, a modularidade (Q) também melhora, passando de 0.18 para 0.28, sugerindo que as comunidades detectadas na rede *backbone* são mais bem definidas. Essa simplificação, tanto em termos de nós quanto de arestas, demonstra a eficácia da técnica de extração de *backbone* em manter a estrutura essencial das comunidades, enquanto reduz o custo computacional. Esses resultados são promissores para a reprodutibilidade de estudos em redes sociais complexas como o Instagram, permitindo análises mais eficientes sem perder a qualidade das informações essenciais.

Analisando os resultados para o Instagram com *RN* para a rede original e o *backbone* apresentados na Tabela 7, observa-se que, para a rede original e amostras de tamanho de 20% dos nós, o NMI já fica bem próximo ao NMI da rede original, com uma redução de aproximadamente 96% das arestas, contribuindo significativamente para a redução da complexidade. No entanto, ao se considerarem os *backbones* das amostras, nota-se uma dificuldade, mesmo com amostras grandes, para atingir valores de NMI próximos daqueles observados no *backbone* originalmente extraído. Isso revela que,

Tabela 7: NMI nas amostras usando o *RN* no Instagram.

Frac. Nós	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Original	.75±.07	.85±.06	.83±.05	.83±.05	.79±.07	.80±.10	.84±.12	.80±.12	.84±.07
Backbone	.49±.08	.58±.04	.62±.04	.67±.03	.69±.04	.66±.07	.73±.04	.77±.03	.77±.03

Tabela 8: NMI nas amostras usando o *FF* no Instagram.

p_f / Frac. Nós	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	.57±.14	.60±.12	.56±.12	.50±.11	.56±.11	.61±.11	.63±.10	.56±.13	.64±.12
0.2	.58±.14	.48±.06	.65±.09	.59±.12	.67±.09	.65±.12	.63±.14	.66±.12	.63±.11
0.3	.55±.12	.59±.12	.68±.07	.56±.12	.59±.09	.59±.14	.60±.11	.61±.12	.61±.14
0.4	.62±.15	.57±.15	.66±.14	.62±.15	.63±.13	.58±.14	.68±.13	.64±.16	.71±.08
0.5	.64±.12	.64±.11	.64±.13	.57±.11	.56±.08	.69±.10	.66±.09	.62±.12	.65±.12
0.6	.59±.15	.64±.12	.52±.13	.57±.10	.65±.11	.55±.15	.57±.10	.64±.12	.71±.09
0.7	.61±.13	.69±.13	.77±.04	.68±.12	.71±.12	.65±.10	.62±.10	.68±.12	.75±.06
0.8	.72±.10	.76±.05	.72±.12	.76±.07	.73±.05	.77±.08	.72±.11	.67±.12	.76±.17
0.9	.70±.12	.83±.08	.81±.06	.73±.08	.74±.12	.75±.08	.78±.06	.78±.07	.76±.09

para esta topologia, a extração de *backbone* pode requerer uma maior atenção para fornecer versões menores e representativas, possivelmente devido à complexidade e densidade das conexões na rede original que são perdidas durante o processo de amostragem e extração de *backbone*. Por exemplo, ao se usar uma amostra de 90% dos nós, o NMI para a rede original é .84±.07, enquanto para o *backbone* é .77±.03 e sem remoção das arestas da rede original, indicando uma perda na qualidade da representação das comunidades no *backbone* e sem ganho computacional.

Os resultados para o Instagram com o método *FF* são apresentados nas Tabelas 8 e 9. Observa-se que este método gera amostras com alta similaridade com as comunidades originais, especialmente para tamanhos de amostra maiores. Para a rede original, a partir de uma amostra de 70%, o NMI se aproxima significativamente do valor obtido para a rede completa. Por exemplo, com uma probabilidade p_f de 0.3, o NMI é .77±.04, próximo ao NMI da rede original, que é .82±.06, e a redução de arestas é de 35%, impondo um custo computacional menor.

Quando se consideram os *backbones* das amostras, observa-se uma maior variabilidade nos valores de NMI. Mesmo com tamanhos de amostra grandes, os valores de NMI não atingem os valores observados no *backbone* originalmente extraído. Isso sugere que, para a topologia do Instagram, a extração de *backbone* a partir das amostras pode não capturar de maneira eficaz a estrutura de comunidade presente no *backbone* original. Além disso, mesmo variando os tamanhos de amostras e parâmetros de probabilidade de queima, os resultados não superam o NMI médio do *backbone* original. Esse comportamento também foi observado na amostragem *RN*. A extração de *backbone*, apesar de simplificar a rede e melhorar a modularidade, pode perder detalhes críticos da estrutura da comunidade, especialmente em redes complexas e densas como o Instagram. Ao analisar amostras de tamanho a partir de 70%, observa-se que a probabilidade de queima de 0.9 proporciona menores reduções na quantidade de arestas, variando de 17% a 0%, exigindo um alto custo computacional para processá-las.

Os resultados para o Instagram com o método *RW* são apresentados na Tabela 10. Observa-se que este método gera amostras com alta similaridade com as comunidades originais, especialmente para tamanhos de amostra maiores. Para amostras de tamanho de 40% dos nós, o NMI se aproxima significativamente do valor obtido para a rede completa e com remoção de 73% na quantidade de arestas.

Tabela 9: NMI nos backbones das amostras usando o *FF* no WhatsApp.

p_f / Frac. Nós	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	.19±.07	.24±.09	.25±.08	.30±.05	.30±.05	.26±.06	.29±.13	.30±.09	.40±.13
0.2	.34±.11	.29±.06	.39±.06	.37±.11	.36±.13	.39±.12	.31±.11	.40±.14	.41±.13
0.3	.36±.05	.42±.10	.43±.14	.40±.06	.43±.10	.40±.15	.37±.09	.45±.09	.42±.10
0.4	.46±.10	.38±.09	.44±.09	.45±.14	.44±.07	.43±.12	.52±.08	.48±.11	.48±.09
0.5	.51±.03	.48±.10	.55±.07	.52±.06	.47±.05	.51±.06	.50±.07	.53±.06	.54±.09
0.6	.57±.05	.56±.03	.56±.04	.53±.04	.56±.04	.55±.07	.57±.05	.61±.07	.60±.04
0.7	.57±.13	.58±.09	.69±.10	.56±.08	.60±.10	.59±.09	.61±.07	.60±.06	.64±.07
0.8	.60±.07	.59±.09	.61±.07	.67±.09	.65±.08	.62±.11	.66±.06	.66±.04	.64±.10
0.9	.71±.04	.72±.03	.72±.05	.74±.04	.71±.06	.73±.05	.73±.03	.73±.05	.71±.01

Tabela 10: NMI nas amostras usando o *RW* no WhatsApp.

Frac. Nós	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Original	.60±.10	.57±.07	.65±.11	.65±.11	.67±.10	.72±.08	.67±.11	.77±.12	.80±.07
Backbone	.50±.08	.59±.05	.63±.05	.67±.03	.69±.05	.67±.08	.74±.03	.76±.04	.77±.03

Além disso, se considerarmos amostras um pouco maiores, podemos obter vantagens quanto a similaridade entre as comunidades, por exemplo, amostras de 60% dos nós apresentam redução em torno de 43% na quantidade de arestas e tem NMI médio de .72±.08 que é mais próximo da rede original, resultando em menor perda na complexidade da estrutura e ainda com algum ganho no custo computacional.

Movendo para os *backbones* das amostras, observa-se uma maior variabilidade nos valores de NMI. Mesmo com tamanhos de amostra grandes, os valores de NMI não atingem os valores observados no *backbone* originalmente extraído. Isso sugere que, para a topologia da rede do Instagram, a extração de *backbone* a partir das amostras pode não capturar de maneira eficaz a estrutura de comunidade presente no *backbone* original. Mesmo variando os tamanhos de amostras, os resultados não superam o NMI médio do *backbone* original. Esse comportamento também foi observado na amostragem *RN*. A extração de *backbone*, apesar de simplificar a rede e melhorar a modularidade, pode perder detalhes críticos da estrutura da comunidade. Considerando amostras de até 60% dos nós, a redução de arestas chega a até 56%, mas, avaliando os valores de NMI, a similaridade é bem baixa com as comunidades do *backbone* original.

Com base nos resultados, podemos concluir que a extração de *backbone* nem sempre favorece a reprodutibilidade da estrutura de comunidade. Considerando a amostragem da rede original, *Random Node (RN)* fornece resultados melhores em termos de NMI. Isso evidencia a seleção de nós, que é menos complexa e requer menos custo em termos de tempo de processamento, consegue garantir a preservação da estrutura de comunidades da rede do Instagram. Já para o *backbone*, os métodos foram ineficazes para representar o *backbone* original, ressaltando a importância de considerar a topologia e a densidade da rede ao selecionar métodos de amostragem, bem como a necessidade de uma avaliação criteriosa da extração de *backbone* para garantir a preservação das estruturas de comunidade essenciais em estudos de reprodutibilidade. Portanto, a escolha do método de amostragem deve ser cuidadosamente ponderada para assegurar que a complexidade e as características inerentes da rede sejam mantidas, permitindo análises representativas.

5.3 Tags Stack Overflow

Os resultados da Tabela 11 destacam a complexidade topológica da rede original do Stack Overflow, com 49.945 nós e 4.147.302 arestas. A alta densidade e o grande número de arestas impõem um custo computacional significativo para a análise. Mesmo após a redução com a extração de *backbone*, a rede permanece considerável, com 41.455 nós e 376.317 arestas. A extração de *backbone* melhora a estrutura de comunidades e o NMI médio.

Tabela 11: Caracterização Topológica e de Comunidades da Rede e do *Backbone* do Stack Overflow.

Rede	# V	# E	$\hat{k}$	CMA	d	# Com.	Q	NMI
Original	49945	4147302	166.07	.63	.0033	14±1.33	.47±.002	.79±.04
Backbone	41455	376317	14.16	.82	.0004	33.3±4.74	.52±.004	.81±.05

Os resultados para o Stack Overflow com o método *RN* para a rede original e o *backbone* são apresentados na Tabela 12. Observa-se um padrão de crescimento do NMI até se aproximar significativamente do valor obtido para a rede completa. Por exemplo, com uma amostra de 90%, o NMI é .71, próximo ao NMI da rede original, .79, com uma redução de 19% na quantidade de arestas. Já com uma amostra de 80%, apesar do NMI ser .63 (cerca de 89% do original), a redução na quantidade de arestas é de 51%, exigindo menos poder computacional para seu processamento.

Focando nos *backbones* das amostras, os valores de NMI também mostram uma melhora gradual à medida que o tamanho da amostra aumenta. No entanto, mesmo com tamanhos de amostra grandes, os valores de NMI não atingem os valores observados no *backbone* originalmente extraído. Isso sugere que, para a topologia da rede, a extração de *backbone* a partir das amostras pode não capturar de maneira eficaz a estrutura de comunidade do *backbone* original. A complexidade e a densidade das conexões na rede original podem ser fatores que influenciam essa dificuldade.

Os resultados da Tabela 13 para o método *FF* indicam uma alta similaridade entre as comunidades amostradas e as comunidades originais. A tabela revela que as amostras geradas, independentemente do tamanho da amostra e da probabilidade de queima, apresentam valores de NMI consistentes e próximos ao da rede original. Por exemplo, uma amostra com tamanho de 30% dos nós e probabilidade de queima de 0.2 apresentou um NMI de .80±.04, o que está muito próximo da rede original (.79±.04) e resultou em uma redução de 34% na quantidade de arestas. Isso sugere que o método é eficaz em preservar a estrutura de comunidade original em amostras pequenas.

Já para os *backbones* das amostras do método *FF*, os resultados apresentados na Tabela 14 mostram que para amostras de tamanhos a partir de 50%, os valores de NMI se aproximam do NMI médio do *backbone* original. Por exemplo, uma amostra com tamanho de 50% dos nós e probabilidade de queima de 0.3 apresentou um NMI de .81±.05 que é equivalente ao NMI do *backbone* original, porém com redução de 23% na quantidade de arestas, o que exige menos custo computacional.

Os resultados da Tabela 15 para o método *Random Walk (RW)* mostram uma alta preservação da estrutura de comunidade original tanto com quanto sem *backbone*. Os resultados mostram que, independentemente do tamanho da amostra, os valores de NMI

Tabela 12: NMI nas amostras usando o *RN* no Stack Overflow.

Frac. Nós	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Original	.24±.04	.31±.06	.35±.04	.41±.06	.49±.07	.54±.08	.59±.06	.63±.05	.71±.05
Backbone	.31±.04	.35±.05	.38±.06	.43±.07	.48±.06	.54±.06	.58±.06	.64±.06	.70±.07

Tabela 13: NMI nas amostras usando o *FF* no Stack Overflow.

p_f / Frac. Nós	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	.76±.02	.75±.02	.76±.02	.75±.01	.76±.02	.76±.02	.76±.03	.75±.02	.76±.02
0.2	.78±.02	.79±.02	.78±.02	.80±.02	.80±.02	.77±.04	.79±.02	.79±.03	.79±.03
0.3	.79±.02	.80±.04	.79±.02	.79±.02	.79±.04	.80±.02	.80±.03	.79±.03	.80±.01
0.4	.78±.04	.79±.03	.78±.03	.78±.04	.77±.03	.78±.03	.79±.04	.78±.03	.77±.04
0.5	.77±.03	.81±.03	.77±.05	.78±.04	.78±.03	.76±.03	.79±.03	.80±.03	.77±.05
0.6	.78±.03	.80±.04	.80±.03	.79±.04	.79±.04	.78±.03	.79±.03	.81±.03	.79±.03
0.7	.79±.04	.80±.04	.79±.04	.77±.04	.78±.04	.77±.03	.78±.03	.80±.03	.77±.03
0.8	.79±.03	.80±.02	.80±.02	.78±.04	.79±.04	.78±.03	.78±.03	.79±.04	.78±.04
0.9	.78±.03	.79±.02	.78±.04	.80±.03	.79±.04	.79±.02	.79±.03	.79±.04	.79±.03

Tabela 14: NMI nos backbones das amostras usando o *FF* no Stack Overflow.

p_f / Frac. Nós	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	.73±.03	.74±.02	.74±.03	.75±.01	.76±.03	.75±.02	.75±.03	.76±.02	.74±.03
0.2	.75±.04	.77±.04	.79±.03	.78±.02	.78±.04	.74±.03	.75±.04	.77±.03	.75±.04
0.3	.77±.04	.78±.03	.78±.04	.79±.03	.78±.04	.79±.03	.80±.03	.78±.02	.79±.03
0.4	.79±.05	.79±.04	.79±.04	.79±.04	.76±.03	.79±.05	.78±.05	.80±.04	.79±.05
0.5	.79±.04	.79±.04	.81±.05	.78±.04	.79±.04	.77±.04	.81±.04	.79±.05	.79±.05
0.6	.79±.04	.79±.05	.80±.03	.78±.04	.80±.04	.78±.04	.78±.03	.79±.04	.83±.02
0.7	.79±.05	.81±.04	.78±.03	.78±.02	.80±.06	.80±.04	.81±.04	.79±.05	.79±.04
0.8	.77±.03	.83±.04	.81±.04	.79±.05	.83±.04	.80±.04	.81±.07	.79±.04	.80±.04
0.9	.80±.04	.81±.04	.82±.03	.80±.04	.79±.04	.80±.04	.79±.04	.80±.04	.79±.04

permanecem consistentemente altos e próximos ao da rede original. Por exemplo, uma amostra com tamanho de 30% dos nós apresentou um NMI de .79±.03 bem próximo ao da rede original de .79±.04, mas com uma redução de 35% na quantidade de arestas. Assim, mesmo com amostras menores, o método se mostra eficaz em manter a integridade das comunidades, evidenciando sua robustez na preservação das características estruturais da rede. Ainda, ao observar os *backbones* das amostras de *RW*, os valores de NMI são altos e próximos da média do NMI do *backbone* original. Por exemplo, uma amostra com tamanho de 40% dos nós apresentou um NMI de .81±.04, comparado ao NMI do *backbone* original de .81±.05, com redução de 23% na quantidade de arestas.

Tabela 15: NMI de amostras *RW* do Stack Overflow.

Frac. Nós	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Original	.74±.02	.78±.03	.79±.03	.80±.02	.79±.03	.80±.03	.79±.03	.79±.03	.78±.02
Backbone	.74±.04	.79±.02	.78±.05	.81±.04	.77±.04	.78±.04	.81±.04	.78±.05	.81±.04

Em suma, os resultados mostram que, para a rede Stack Overflow, a extração de *backbone* pode exigir amostras maiores para favorecer a reprodutibilidade da estrutura de comunidade. Os métodos de amostragem baseados na exploração de nós, como *FF* e *RW*, fornecem resultados melhores em termos de NMI quando comparados às amostras de *backbone*. Assim, apesar de requererem tempo de processamento mais alto, esses métodos foram capazes de assegurar a fidelidade estrutural necessária para a análise precisa das comunidades da rede.

6 Conclusões e Trabalhos Futuros

A importância deste estudo reside na crescente necessidade de pesquisas que busquem formas de reduzir o tamanho das redes, preservando sua topologia e as propriedades dos nós. A simplificação de redes, minimizando a perda de informações, é crucial para obter representações mais manejáveis e, ao mesmo tempo, representativas dessas redes. No entanto, pouco foco é dado a uma tarefa importante no contexto de redes, a detecção de comunidades. Neste trabalho, demos um primeiro passo nessa direção, fornecendo uma metodologia para analisar a preservação da estrutura de comunidades em redes sociais utilizando diferentes técnicas de amostragem e extração de *backbone*.

À luz da Teoria de redes sociais, nossos resultados mostraram que é possível reduzir significativamente o tamanho das redes com perdas mínimas ou até sem perdas substanciais na detecção de comunidades e obter menor custo computacional devido à redução da quantidade de arestas. Observamos a importância de considerar diferentes métodos de amostragem e suas parametrizações para garantir a reprodutibilidade de estudos em redes sociais. Além das contribuições metodológicas, os resultados obtidos possuem implicações práticas significativas, permitindo que pesquisadores com recursos computacionais limitados realizem análises escaláveis e investiguem fenômenos sociais, científicos e tecnológicos que, de outra forma, seriam inviáveis em redes de grande escala. Apesar dos avanços apresentados, os métodos de amostragem e extração de *backbone* podem introduzir vieses que afetam a reprodutibilidade, especialmente em redes com topologias específicas. Sua eficácia varia conforme a estrutura da rede, sendo mais adequada para redes sociais do que para outras, como as de infraestrutura [13]. Para trabalhos futuros, investigaremos como as características topológicas influenciam a qualidade da amostra, considerando os benefícios em custo computacional e as limitações em diferentes tipos de redes.

Agradecimentos

Os autores agradecem à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Referências

[1] Nesreen K Ahmed, Jennifer Neville, and Ramana Kompella. 2013. Network sampling: From static to streaming graphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 8, 2 (2013), 1–56.

[2] Yasir Arfat, Sugimiyanto Suma, Rashid Mehmood, and Aiiad Albeshri. 2020. Parallel shortest path big data graph computations of US road network using apache spark: survey, architecture, and evaluation. *Smart Infrastructure and Applications: Foundations for Smarter Cities and Societies* (2020), 185–214.

[3] Albert-László Barabási et al. 2016. *Network science.* Cambridge university press.

[4] Punam Bedi and Chhavi Sharma. 2016. Community detection in social networks. *Wiley interdisciplinary reviews: Data mining and knowledge discovery* (2016).

[5] Austin R Benson, Rediet Abebe, Michael T Schaub, Ali Jadbabaie, and Jon Kleinberg. 2018. Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences* (2018), 11221–11230.

[6] Austin R Benson, Ravi Kumar, and Andrew Tomkins. 2018. Sequences of sets. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

[7] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[8] Pham Minh Chuan, Le Hoang Son, Mumtaz Ali, Tran Dinh Khang, Le Thanh Huong, and Nilanjan Dey. 2018. Link prediction in co-authorship networks based on hybrid content similarity metric. *Applied Intelligence* 48 (2018), 2470–2486.

[9] Michele Coscia. 2021. Noise Corrected Sampling of Online Social Networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 2 (2021), 1–21.

[10] Michele Coscia and Luca Rossi. 2019. The impact of projection and backboning on network topologies. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 286–293.

[11] Marc Cubrich, Rachel T. King, Derek L. Mracek, Jamie M.G. Strong, Kristen Hassenkamp, Daly Vaughn, and Nikki M. Dudley. 2021. Examining the criterion-related validity evidence of LinkedIn profile elements in an applied sample. *Comput. Hum. Behav.* (2021).

[12] Jose Martins da Rosa, Renan Saldanha Linhares, Carlos Henrique Gomes Ferreira, Gabriel P. Nobre, Fabricio Murai, and Jussara M. Almeida. 2022. Uncovering Discussion Groups on Claims of Election Fraud from Twitter. In *Proc. of Social Informatics: 13th International Conference.* https://doi.org/10.1007/978-3-031-19097-1_20

[13] Liang Dai, Ben Derudder, and Xingjian Liu. 2018. Transport network backbone extraction: A comparison of techniques. *Journal of Transport Geography* (2018).

[14] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. 2005. Comparing community structure identification. *Journal of statistical mechanics: Theory and experiment* 2005, 09 (2005), P09008.

[15] Zahir Edrees. 2020. Network Analysis of the Stack Overflow Tags. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 44 (2020), 195–202.

[16] Meihua Fan, Shudong Li, Weihong Han, Xiaobo Wu, Zhaoquan Gu, and Zhihong Tian. 2020. A novel malware detection framework based on weighted heterograph. In *Proceedings of the 2020 International Conference on Cyberspace Innovation of Advanced Technologies*. 39–43.

[17] Carlos HG Ferreira, Fabricio Murai, Ana PC Silva, Jussara M Almeida, Martino Trevisan, Luca Vassio, Marco Mellia, and Idilio Drago. 2021. On the dynamics of political discussions on instagram: A network perspective. *Online Social Networks and Media* 25 (2021), 100155.

[18] Santo Fortunato and Darko Hric. 2016. Community detection in networks: A user guide. *Physics reports* 659 (2016), 1–44.

[19] Xiang Fu, Shangdi Yu, and Austin R Benson. 2020. Modelling and analysis of tagging networks in Stack Exchange communities. *Journal of Complex Networks* 8, 5 (2020), cnz045.

[20] Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. 2010. Walking in facebook: A case study of unbiased sampling of osns. In *2010 Proceedings IEEE Infocom*. Ieee, 1–9.

[21] Carlos Henrique Gomes Ferreira, Fabricio Murai, Ana PC Silva, Martino Trevisan, Luca Vassio, Idilio Drago, Marco Mellia, and Jussara M Almeida. 2022. On network backbone extraction for modeling online collective behavior. *Plos one* 17, 9 (2022), e0274218.

[22] Douglas D Heckathorn and Christopher J Cameron. 2017. Network sampling: From snowball and multiplicity to respondent-driven sampling. *Annual review of sociology* 43 (2017), 101–119.

[23] Pili Hu and Wing Cheong Lau. 2013. A survey and taxonomy of graph sampling. *arXiv preprint arXiv:1308.5865* (2013).

[24] Luke Hutton and Tristan Henderson. 2015. Toward reproducibility in online social network research. *IEEE Transactions on Emerging Topics in Computing* (2015).

[25] Jeancarlo C Leao, Michele A Brandao, Pedro OS Vaz de Melo, and Alberto HF Laender. 2017. Classificação de relações sociais para melhorar a detecção de comunidades. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC.

[26] Jure Leskovec and Christos Faloutsos. 2006. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 631–636.

[27] Renan S. Linhares, José M. Rosa, Carlos H. G. Ferreira, Fabricio Murai, Gabriel Nobre, and Jussara Almeida. 2022. Uncovering Coordinated Communities on Twitter During the 2020 U.S. Election. In *IEEE/ACM international conference on advances in social networks analysis and mining*.

[28] László Lovász. 1993. Random walks on graphs. *Combinatorics, Paul erdos is eighty* 2, 1-46 (1993), 4.

[29] Anna May, Johannes Wachs, and Anikó Hannák. 2019. Gender differences in participation and reward on Stack Overflow. *Empirical Software Engineering* 24 (2019), 1997–2019.

[30] Cong Mu, Youngser Park, and Carey E Priebe. 2023. Dynamic network sampling for community detection. *Applied Network Science* 8, 1 (2023), 5.

[31] Ryan Murtfeldt, Naomi Alterman, Ihsan Kahveci, and Jevin D West. 2024. RIP Twitter API: A eulogy to its vast research contributions. *arXiv preprint arXiv:2404.07340* (2024).

[32] Zachary P Neal. 2022. backbone: An R package to extract network backbones. *PloS one* 17, 5 (2022), e0269137.

[33] Mark Newman. 2018. *Networks*. Oxford university press.

[34] Mark EJ Newman and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical review E* 69 (2004), 026113.

[35] Gabriel Peres Nobre, Carlos Henrique Gomes Ferreira, and Jussara Marques Almeida. 2020. Beyond groups: Uncovering dynamic communities on the whatsapp network of information dissemination. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy*. Springer, 252–266.

[36] Randal S Olson and Zachary P Neal. 2015. Navigating the massive world of reddit: Using backbone networks to map user interests in social media. *PeerJ Computer Science* 1 (2015), e4.

[37] Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. 2021. Uncovering Coordinated Networks on Social Media: Methods and Case Studies. In *International Conference on Web and Social Media*.

[38] Anne Plant and Robert Hanisch. 2020. Reproducibility in science: A metrology perspective. *Harvard Data Science Review* 2, 4 (2020).

[39] Pascal Pons and Matthieu Latapy. 2005. Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005: 20th International Symposium*.

[40] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society* 7, 2 (2021), 20563051211019004.

[41] Filippo Radicchi, José J. Ramasco, and Santo Fortunato. 2011. Information filtering in complex weighted networks. *Phys. Rev. E* (2011).

[42] Katerina Rigana, Ernst-Jan Camiel Wit, and Samantha Cook. 2023. A new way of measuring effects of financial crisis on contagion in currency markets. *International Review of Financial Analysis* 90 (2023), 102764.

[43] Hamid Roghani, Asgarali Bouyer, and Esmaeil Nourani. 2021. PLDLS: A novel parallel label diffusion and label Selection-based community detection algorithm based on Spark in social networks. *Expert Systems with Applications* (2021).

[44] Giulio Rossetti and Rémy Cazabet. 2018. Community discovery in dynamic networks: a survey. *Comput. Surveys* 51 (2018), 35.

[45] Benedek Rozemberczki, Oliver Kiss, and Rik Sarkar. 2020. Little Ball of Fur: A Python Library for Graph Sampling. In *ACM International Conference on Information and Knowledge Management*. ACM.

[46] S Haleh S. Dizaji, Joze M Rozanec, Reza Farahani, Dumitru Roman, and Radu Prodan. 2024. An Extensive Characterization of Graph Sampling Algorithms. In *Companion of the 15th ACM/SPEC International Conference on Performance Engineering*. 135–140.

[47] M Ángeles Serrano, Marián Boguná, and Alessandro Vespignani. 2009. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences* (2009).

[48] Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5, 1 (2001), 3–55.

[49] Francesca Soro, Mauro Allegretta, Marco Mellia, Idilio Drago, and Leandro M Bertholdo. 2020. Sensing the noise: Uncovering communities in darknet traffic. In *2020 Mediterranean Communication and Computer Networking Conference (MedComNet)*. IEEE, 1–8.

[50] Statista. 2024. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025. Retrieved June 18, 2024 from https://www.statista.com/statistics/871513/worldwide-data-created/

[51] Michael PH Stumpf, Carsten Wiuf, and Robert M May. 2005. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences* 102, 12 (2005), 4221–4224.

[52] Cong Tran, Won-Yong Shin, and Andreas Spitz. 2021. Community detection in partially observable social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16, 2 (2021), 1–24.

[53] Alexander van der Grinten, Eugenio Angriman, and Henning Meyerhenke. 2020. Scaling up network centrality computations–A brief overview. *it-Information Technology* 62, 3-4 (2020), 189–204.

[54] Otavio R Venâncio, Carlos HG Ferreira, Jussara M Almeida, and Ana Paula C da Silva. 2024. Unraveling User Coordination on Telegram: A Comprehensive Analysis of Political Mobilization during the 2022 Brazilian Presidential Election. In *International AAAI Conference on Web and Social Media*.

[55] Michael S Vitevitch and Mary Sale. 2023. Identifying the phonological backbone in the mental lexicon. *Plos one* 18, 6 (2023), e0287197.

[56] Dong Wang, Zhenyu Li, and Gaogang Xie. 2011. Towards unbiased sampling of online social networks. In *2011 IEEE International Conference on Communications (ICC)*. IEEE, 1–5.

[57] Yi Wang. 2018. Understanding the reputation differences between women and men on stack overflow. In *2018 25th Asia-Pacific Software Engineering Conference (APSEC)*. IEEE, 436–444.

[58] Song Xinchao and Yishuang Geng. 2014. Distributed community detection optimization algorithm for complex networks. *Journal of Networks* 9, 10 (2014), 2758.

[59] Jianpeng Zhang, Hongchang Chen, Dingjiu Yu, Yulong Pei, and Yingjun Deng. 2023. Cluster-preserving sampling algorithm for large-scale graphs. *Science China Information Sciences* 66, 1 (2023), 112103.

[60] Junzhou Zhao, Pinghui Wang, John CS Lui, Don Towsley, and Xiaohong Guan. 2019. Sampling online social networks by random walk with indirect jumps. *Data Minifng and Knowledge Discovery* 33 (2019), 24–57.