Interseção de Vizinhança em Grafos via Amostragem de P_3

Vinícius M. Ribeiro[1], **André L. Vignatti**[1]

[1]Departamento de Informática – Universidade Federal do Paraná (UFPR)
Curitiba – PR – Brasil

{vmr20,vignatti}@inf.ufpr.br

Abstract. *Neighborhood intersection is a fundamental metric in social network analysis and data mining, and plays a central role in the computation of similarity metrics and measures. In this paper, we propose an efficient randomized algorithm, using P_3 sampling, to compute the neighborhood intersection for all combinations of pairs of vertices in a graph. With probability at least $1 - \delta$, we ensure that all approximations of the algorithm are at most ϵ away from their true value. We apply techniques from computational learning theory to obtain a sample size independent of any quantitative property of the graph.*

Resumo. *A interseção de vizinhança é uma métrica fundamental em análise de redes sociais e mineração de dados, e desempenha um papel central no cálculo de métricas e medidas de similaridade. Nesse artigo, propomos um algoritmo aleatorizado eficiente, utilizando amostragem de P_3, para calcular a interseção de vizinhança para todas as combinações de pares de vértices em um grafo. Com probabilidade de pelo menos $1 - \delta$, garantimos que todas as aproximações do algoritmo estão a uma distância de no máximo ϵ de seu valor real. Aplicamos técnicas da teoria de aprendizado computacional para obter um tamanho de amostra independente de qualquer propriedade quantitativa do grafo.*

1. Introdução

A *interseção de vizinhança* de dois vértices em um grafo é definida como $|N(u) \cap N(v)|$, onde $N(u)$ e $N(v)$ são as vizinhanças dos vértices u e v, respectivamente. Ela desempenha um papel central no cálculo de métricas e medidas de similaridade [Easley and Kleinberg 2010, Newman 2010, Menczer et al. 2020]. Muitas métricas estabelecidas (como similaridade de cosseno, sobreposição de vizinhança e equivalência estrutural) são baseadas em versões normalizadas da interseção de vizinhança. Por isso, versões normalizadas podem ser de maior interesse do que a métrica não normalizada. A normalização é útil, pois fornece valores relativos ao grafo, que podem ser mais informativos do que medidas absolutas [Ribeiro and Vignatti 2025]. Neste artigo, introduzimos a interseção da vizinhança P_3-normalizada, uma nova normalização para a interseção da vizinhança. A desnormalização dessa métrica, que permite obter resultados diretos para a métrica original, pode ser realizada com técnicas de Ribeiro e Vignatti [Ribeiro and Vignatti 2025], mas não será tratada aqui devido a limitações de espaço. Apresentaremos um algoritmo para determinar, para todos os pares de vértices de um grafo, o tamanho da interseção de vizinhança normalizada, baseado na amostragem de caminhos P_3. Ao contrário das abordagens tradicionais que utilizam amostragem de estruturas mais simples, como vértices e arestas [Ribeiro and Vignatti 2025], nossa técnica explora a amostragem de uma estrutura mais complexa, resultando em um algoritmo mais eficiente. Desenvolvemos um método de amostragem uniforme e eficiente

de P_3, um desafio não trivial em comparação com a amostragem de vértices ou arestas, que pode ser de interesse geral. Além disso, aplicamos técnicas avançadas de amostragem da teoria do aprendizado computacional para determinar limitantes rigorosos para o número de amostras necessárias, garantindo parâmetros de erro e confiança desejados. Embora a computação exata de interseção de vizinhança seja pouco estudada, existem métodos aproximados, como os de Besta et al. [Besta et al. 2021, Besta et al. 2022] para interseções únicas (apenas um par de vértices). Além disso, Ribeiro e Vignatti [Ribeiro and Vignatti 2025] abordam o problema de interseção de vizinhança de todos os pares, alcançando tempo $O(\Delta \log \Delta + |E|)$ em sua melhor estratégia, onde Δ é o grau máximo do grafo $G = (V, E)$. No mesmo cenário, nossa abordagem de amostragem de P_3 apresenta um tempo de execução de $O(|E|)$, demonstrando maior eficiência teórica e justificando a relevância do presente trabalho.

2. Preliminares

Seja $G = (V, E)$ um grafo não direcionado. Um P_3 é definido como sendo três vértices $\{u, v, w\}$ que possuem as arestas $\{u, v\}$ e $\{v, w\}$. $\mathbb{P}_3$ é o conjunto de todos os P_3 de G.

Definição 1. A *interseção da vizinhança* $i(u, v)$ de dois vértices u e v é a quantidade de vizinhos em comum de u e v, e.g. $i(u, v) = |N(u) \cap N(v)|$.

Conforme explicado na Seção 1, a normalização dos valores $i(u, v)$ não apenas gera resultados úteis, mas também enriquece o seu significado. Neste trabalho, focamos no cálculo de uma versão normalizada específica, apresentada na Definição 2.

Definição 2. A *interseção da vizinhança P_3-normalizada* $i_{P3}(u, v)$ do par de vértices u, v é dada por $i_{\text{P3}}(u, v) = \frac{i(u,v)}{|\mathbb{P}_3|}$.

Os valores $i_{\text{P3}}(u, v)$ representam o quão expressiva é a interseção da vizinhança de um par de vértices u, v, independente de qualquer propriedade do grafo. Se $i_{\text{P3}}(u, v) = 1$, então u e v compartilham a maior quantidade possível de vizinhos. Se $i_{\text{P3}}(u, v) = 0$, então u e v não compartilham nenhum vizinho.

Quando desejamos estimar múltiplos valores simultaneamente, a estratégia comum envolve a obtenção de limitantes individuais usando desigualdades probabilísticas clássicas, como as de Chernoff e Hoeffding, e, em seguida, a combinação desses limitantes individuais para obter um limitante global desejado através do limitante da união. Uma desvantagem dessa abordagem é que o limite da união leva a um tamanho de amostra que depende da quantidade de valores que se deseja estimar. A dimensão Vapnik-Chervonenkis (VC), proveniente da teoria de aprendizagem computacional [Shalev-Shwartz and Ben-David 2014], oferece um método mais refinado, limitando o tamanho da amostra com base na "complexidade" do conjunto de valores, em vez de sua cardinalidade. A seguir, apresentamos definições e teoremas a respeito desse assunto.

Um *espaço de intervalos* $(X, \mathcal{R})$ consiste em um conjunto X e uma família $\mathcal{R}$ de subconjuntos de X. A *projeção* de $A \subseteq X$ em $\mathcal{R}$ é $P_{\mathcal{R}}(A) = \{A \cap R : R \in \mathcal{R}\}$. Um conjunto A é *estilhaçado* por $\mathcal{R}$ se $P_{\mathcal{R}}(A) = 2^A$. A *dimensão VC* de um espaço de intervalo $(X, \mathcal{R})$, denotado $d_{\text{VC}}(\mathcal{R})$, é o tamanho do maior conjunto $A \subseteq X$ estilhaçado por $\mathcal{R}$. Para uma explicação mais aprofundada da dimensão VC, veja [Shalev-Shwartz and Ben-David 2014].

Definição 3. Seja $(X, \mathcal{R})$ um espaço de intervalo com uma distribuição de probabilidade π sobre X, $p_R = \Pr_\pi(R)$ a probabilidade de um intervalo $R \in \mathcal{R}$, e $\hat{p_R}$ a frequência

relativa de R com base em uma amostra S. Para $\epsilon \in (0,1)$, S é uma *ϵ-aproximação* para $(X, \mathcal{R})$ se $|\hat{p_R} - p_R| \leq \epsilon, \forall R \in \mathcal{R}$.

Teorema 1 ([Har-Peled and Sharir 2011], Teo. 2.12)**.** *Seja $(X, \mathcal{R})$ um espaço de intervalo com dimensão VC $d_{VC}(\mathcal{R}) \leq d$, e seja π uma distribuição de probabilidade em X. Para $\epsilon, \delta \in (0,1)$, e uma amostra S de tamanho m extraída de π, com probabilidade pelo menos $1-\delta$, S é uma ϵ-aproximação de $(X, \mathcal{R})$ se $m \geq \frac{c}{\epsilon^2}\left(d + \ln\frac{1}{\delta}\right)$, onde c é uma constante positiva.*

Por limitações de espaço, as provas dos teoremas serão omitidas neste trabalho e apresentadas em sua respectiva versão estendida.

3. Estimativa Usando Amostragem de P_3

Esta seção apresenta a nossa estratégia para estimar as interseções normalizadas $i_{\text{P3}}(u, v)$.

3.1. Espaço de Intervalos e Resultados de Dimensão VC

A quantidade de amostras necessárias ao algoritmo baseia-se no resultado do Teorema 1, que, por sua vez, utiliza os valores da dimensão VC de um espaço de intervalos definidos para o problema. Abordaremos esse assunto a seguir.

Seja $\mathbb{P}_3$ o espaço amostral e $E_{P3}(u, v)$ o evento em que uma amostra P_3 pertence à interseção das vizinhanças de u e v. Temos que

$$\Pr(E_{P3}(u, v)) = \frac{|N(u) \cap N(v)|}{|\mathbb{P}_3|} = i_{\text{P3}}(u, v),$$

onde a última equação segue da Definição 2. Assim, ao amostrar caminhos P_3, aproximamos $i_{\text{P3}}(u, v)$ estimando a probabilidade de $E_{\text{P3}}(u, v)$. Seja $(\mathbb{P}_3, \mathcal{R}_{P3})$ um espaço de intervalos, onde $\mathcal{R}_{P3}$ é o conjunto de todos os eventos $E_{P3}(u, v)$ para $u, v \in V$, i.e., $\mathcal{R}_{P3} = \{E_{P3}(u, v) \mid u, v \in V\}$. O Teorema 2 limita a dimensão VC de $(\mathbb{P}_3, \mathcal{R}_{P3})$.

Teorema 2. *A dimensão VC de $(\mathbb{P}_3, \mathcal{R}_{P3})$ é $d_{VC}(\mathcal{R}_{P3}) = 1$ se $|\mathbb{P}_3| \geq 1$, e 0 caso contrário.*

3.2. Algoritmo

Como discutido na seção anterior, para calcular a interseção da vizinhança de u e v $i_{\text{P3}}(u, v)$, podemos estimar a probabilidade do evento $E_{P3}(u, v)$. Utilizamos a frequência relativa de $E_{P3}(u, v)$ para definir o nosso estimador $\hat{i}_{\text{P3}}(u, v)$, i.e.,

$$\hat{i}_{\text{P3}}(u, v) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{E_{P3}(u,v)}(s_i),$$

onde $\mathbb{1}_{E_{P3}(u,v)}(s_i)$ é uma variável indicadora que assume 1 se $s_i \in E_{P3}(u, v)$ e 0 caso contrário. Note que $\hat{i}_{\text{P3}}(u, v)$ é um estimador não enviesado para $i_{\text{P3}}(u, v)$, ou seja, $\mathbb{E}[\hat{i}_{\text{P3}}(u, v)] = i_{\text{P3}}(u, v)$. Tendo definido $i_{\text{P3}}(u, v)$ e $\hat{i}_{\text{P3}}(u, v)$ como a probabilidade de um evento e sua frequência relativa, respectivamente, podemos construir uma ϵ-aproximação (Definição 3) para $(\mathbb{P}_3, \mathcal{R}_{P3})$. Formalmente, nosso algoritmo satisfaz,

$$\Pr\left(\forall u, v \in V, \left|i_{\text{P3}}(u, v) - \hat{i}_{\text{P3}}(u, v)\right| \leq \epsilon\right) \geq 1 - \delta.$$

Ou seja, com probabilidade pelo menos $1 - \delta$, o Algoritmo 1 garante que as estimativas para *todos* os pares de vértices estejam dentro de um erro máximo de ϵ em relação aos seus valores reais.

Algoritmo 1: INTERSEÇÃOVIZINHANÇAALEATORIZADO(G, ϵ, δ)

Entrada: Grafo $G = (V, E)$, precisão ϵ, confiança $1 - \delta$
Saída : Estimativa $\hat{i}_{\text{P3}}(u, v)$ para todo par $u, v \in V$

$m \leftarrow \left\lceil \frac{c}{\epsilon^2} \left(1 + \ln \frac{1}{\delta}\right) \right\rceil$
for $i \leftarrow 1$ **to** m **do**
  $\{u, x, v\} \leftarrow$ AMOSTRARP3(G)
  $\hat{i}_{\text{P3}}(u, v) \leftarrow \hat{i}_{\text{P3}}(u, v) + \frac{1}{m}$
return $\hat{i}_{\text{P3}}(u, v)$ para todos os pares $u, v \in V$

Teorema 3. *Dado um grafo $G = (V, E)$, parâmetros ϵ e δ, as estimativas $\hat{i}_{P3}(u, v)$ retornadas pelo Algoritmo 1 satisfazem* $\Pr\left(\forall u, v \in V, \left| i_{P3}(u, v) - \hat{i}_{P3}(u, v) \right| \leq \epsilon\right) \geq 1 - \delta$.

O tempo de execução do Algoritmo 1 depende do tempo de execução da função AMOSTRARP3, que será apresentado na Seção 3.3.

3.3. Amostragem Uniforme de P_3

Nesta seção, descrevemos um método para amostrar uniformemente um P_3 de um grafo G. Para cada aresta $e = \{u, v\} \in E$, definimos $p_e = \frac{d_u + d_v - 2}{2|\mathbb{P}_3|}$, onde d_u e d_v são os graus dos vértices u e v, respectivamente. A amostragem com pesos p_e pode ser feita usando o método Alias [Walker 1974] em tempo $O(|E|)$ para pré-processar os dados do grafo, e $O(1)$ para fazer cada amostragem. O Algoritmo 2 descreve o processo.

Algoritmo 2: AMOSTRARP3(G)

Entrada: Grafo $G = (V, E)$
Saída : Um P_3 amostrado uniformemente de $\mathbb{P}_3$

1. Amostre uma aresta $e = \{u, x\}$ com probabilidade p_e.
2. Amostre um vértice v de $N(u) \cup N(x) \setminus \{u, x\}$ uniformemente.
3. Retorne o caminho $\{u, x, v\}$.

Para definir os valores p_e no Passo 1, é necessário saber $|\mathbb{P}_3|$. O Lema 4 fornece um resultado útil para esse propósito.

Lema 4. *O número total de P_3 no grafo G é dado por* $|\mathbb{P}_3| = \sum_{\{u,v\} \in E} \frac{d_u + d_v - 2}{2}$.

O Teorema 5 mostra a corretude e o tempo de execução do Algoritmo 2. O tempo de execução, em particular, é baseado no método Alias e no Lema 4.

Teorema 5. *O Algoritmo 2 é de tempo $O(|E|)$ e amostra uniformemente um P_3 de $\mathbb{P}_3$.*

Finalmente, usando os resultados apresentados, podemos enunciar o Teorema 6.

Teorema 6. *O tempo de execução do Algoritmo 1 é $O(|E|)$.*

4. Considerações Finais

Propomos um algoritmo que computa a interseção normalizada entre todos os pares de vértices com erro ϵ e probabilidade $1 - \delta$. Seu tempo de execução é $O(|E|)$, sendo mais eficiente do que outros no mesmo cenário. Ele utiliza amostragem de P_3, que acaba sendo o gargalo do tempo de execução, e métodos mais eficientes de amostragem, como MCMC (Monte Carlo via Cadeias de Markov), podem ser explorados no futuro.

Referências

Besta, M., Kanakagiri, R., Kwasniewski, G., Ausavarungnirun, R., Beránek, J., Kanellopoulos, K., Janda, K., Vonarburg-Shmaria, Z., Gianinazzi, L., Stefan, I., Luna, J. G., Golinowski, J., Copik, M., Kapp-Schwoerer, L., Di Girolamo, S., Blach, N., Konieczny, M., Mutlu, O., and Hoefler, T. (2021). Sisa: Set-centric instruction set architecture for graph mining on processing-in-memory systems. In *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*, page 282–297, New York, NY, USA. Association for Computing Machinery.

Besta, M., Miglioli, C., Labini, P. S., Tětek, J., Iff, P., Kanakagiri, R., Ashkboos, S., Janda, K., Podstawski, M., Kwaśniewski, G., Gleinig, N., Vella, F., Mutlu, O., and Hoefler, T. (2022). Probgraph: high-performance and high-accuracy graph mining with probabilistic set representations. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, Dallas, Texas. IEEE Press.

Easley, D. and Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, USA.

Har-Peled, S. and Sharir, M. (2011). Relative (p, ϵ)-approximations in geometry. *Discrete Comput. Geom.*, 45(3):462–496.

Menczer, F., Fortunato, S., and Davis, C. (2020). *A First Course in Network Science*. Cambridge University Press, UK.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press, Inc., UK.

Ribeiro, V. M. and Vignatti, A. L. (2025). Efficient approximations of neighborhood intersection in large graphs via sampling. Technical report, Federal University of Paraná.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, UK.

Walker, A. (1974). New fast method for generating discrete random numbers with arbitrary frequency distributions. *Electronics Letters*, 10:127–128.