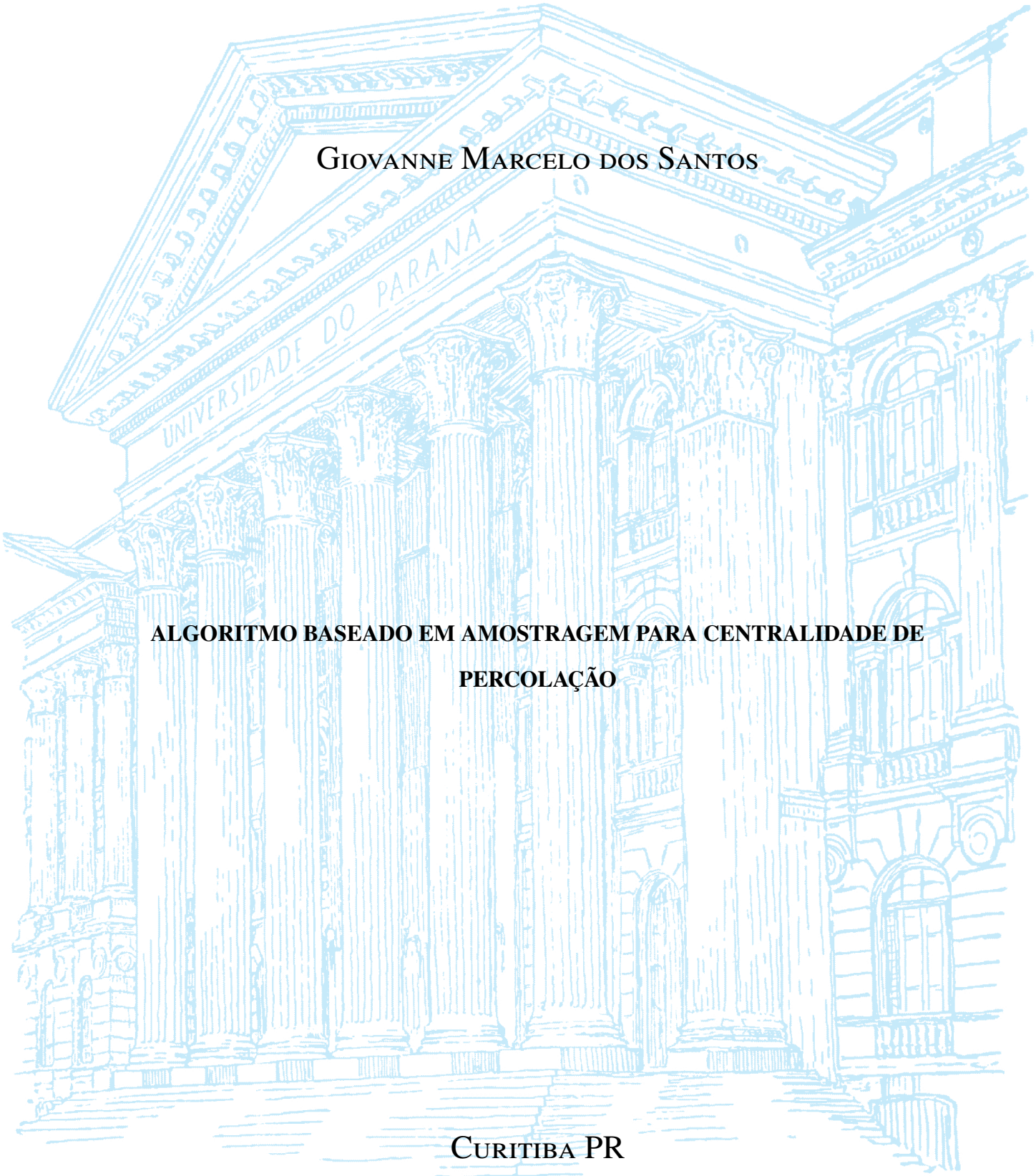
Universidade Federal do Paraná

Giovanne Marcelo dos Santos



ALGORITMO BASEADO EM AMOSTRAGEM PARA CENTRALIDADE DE PERCOLAÇÃO

Curitiba PR
2018

Giovanne Marcelo dos Santos

ALGORITMO BASEADO EM AMOSTRAGEM PARA CENTRALIDADE DE PERCOLAÇÃO

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Prof. Dr. André Luís Vignatti.

Curitiba PR
2018

Resumo

As redes são ferramentas para o estudo de sistemas complexos. Com elas, pesquisadores podem extrair propriedades da estrutura dos sistemas e compreendê-los melhor. Uma das propriedades que pode ser extraída é a *centralidade*. Essa medida quantifica a importância relativa das entidades ou conexões de uma rede. Como o conceito de importância é relativo ao contexto em que está sendo aplicada, existem diferentes medidas de centralidade, como por exemplo a *centralidade de grau*, a *centralidade de intermediação* e a *centralidade de percolação*. O conceito de centralidade de percolação utiliza a Teoria da Percolação como base para determinar a importância de uma entidade durante uma infestação na rede. O algoritmo conhecido para computar essa medida em uma rede com n entidades possui complexidade de tempo $\Theta(n^3)$, podendo ser reduzida para $\Theta(nm)$ se não for considerada a entidade de destino no cálculo. Com base nisso, este trabalho apresenta um algoritmo aleatorizado baseado em amostragem para estimar a centralidade de percolação, que considera as entidades fonte e destino no cálculo. O algoritmo proposto executa em tempo $O(max(n^2, (n+m)\frac{1}{\epsilon^2}\ln\frac{1}{\delta}))$ para redes sem peso e $O(max(n^2, (m + n\log n)\frac{1}{\epsilon^2}\ln\frac{1}{\delta})$ para redes com peso, apresentando um erro de no máximo ϵ com probabilidade $1-\delta$. Na análise do número de amostras necessárias para o algoritmo, são utilizados os resultados da Teoria da Dimensão Vapnik–Chervonenkis e ϵ-amostra.

Palavras-chave: Algoritmo Aleatorizado, Amostragem, Centralidade, Dimensão Vapnik-Chervonenkis, ϵ-amostra, Complexidade de Amostra.

Abstract

Networks are tools to study complex systems. With them, researchers can extract structural properties of the systems and understand them better. One of the properties that can be extracted is the *centrality*. This measure quantifies the relative importance of the entities or links in a network. Since the concept of importance is relative to the applied context, there are different centrality measures, such as *degree centrality, betweenness centrality* and *percolation centrality*. The latter uses the Percolation Theory as the basis for determining the importance of an entity during an infestation in the network. The known algorithm to compute this measure in a network with n entities has $\Theta(n^3)$ time complexity, which can be reduced to $\Theta(nm)$ if the target entity is not considered in the calculation. Based on this, the present work shows a randomized algorithm based on sampling to estimate the percolation centrality, which considers the source and target entities in the calculation. The proposed algorithm runs in $O(max(n^2, (n+m)\frac{1}{\epsilon^2}\ln\frac{1}{\delta}))$ time for unweighted networks and in $O(max(n^2, (m + n\log n)\frac{1}{\epsilon^2}\ln\frac{1}{\delta})$ time for weighted networks. In the analisys of the required sample size for the algorithm, the Vapnik-Chervonenkis Dimension Theory and ϵ-sample are used.

Keywords: Randomized Algorithms, Sampling, Centrality, Vapnik-Chervonenkis Dimension, ϵ-sample, Sample Complexity.

Lista de Figuras

Sumário

1 Introdução

Uma rede é uma maneira de representar padrões de conexões entre entidades de um sistema, como visto em Newman (2010). A internet, a rede de telefonia, a rede de rodovias que interligam cidades, e as redes sociais são exemplos de redes. Esses exemplos representam sistemas que são de interesse para pesquisadores, que, extraindo propriedades da estrutura dessas redes, podem ter uma melhor compreensão do comportamento desses sistemas. Uma das propriedades utilizadas no estudo das redes é a *centralidade*. Essa medida quantifica a importância relativa das entidades ou conexões que compõem o sistema.

Como o conceito de importância é dependente do contexto em que é aplicado, existem diferentes medidas de centralidade, como a *centralidade de grau* e a *centralidade de intermediação*. Então, tendo em vista a necessidade de uma medida de centralidade para casos em que o sistema está passando por um processo de infestação (e.g. transmissão de uma doença entre cidades ou divulgação de notícias falsas em redes sociais), Piraveenan et al. (2013) propõem a *centralidade de percolação*. Essa medida proposta tem como base a centralidade de intermediação e trabalhos como Newman (2002) e Sander et al. (2002), que modelam epidemias como aplicações do processo de percolação em redes. Além disso, Piraveenan et al. (2013) também demonstram que a complexidade de tempo do algoritmo para a centralidade de percolação é da mesma ordem que a do algoritmo para o cálculo da centralidade de intermediação.

Existem algoritmos exatos para a centralidade de intermediação que executam em tempo polinomial, como o apresentado por Brandes (2001). Porém, segundo Riondato e Kornaropoulos (2016), como as redes podem ter milhões ou bilhões de entidades e conexões, esses algoritmos não executam bem na prática. Sendo assim, com o objetivo de diminuir a complexidade de tempo do algoritmo e permitir ao usuário a escolha entre tempo de execução menor e maior precisão na resposta, Riondato e Kornaropoulos (2016) apresentam um algoritmo aleatorizado para aproximar a centralidade de intermediação, baseado em amostragem de caminhos mínimos. Para a análise do número de caminhos mínimos necessários, foram utilizados os resultados da Teoria da Dimensão Vapnik-Chervonenkis (VC).

Com base nisso, este trabalho tem como objetivo propor um algoritmo aleatorizado para estimar a centralidade de percolação de uma entidade na rede, utilizando amostras de caminhos mínimos. Este trabalho está restringido à centralidade de percolação que considera que os vértices fonte e destino dos caminhos mínimos estejam percolados. Além disso, será mostrado como reformular o problema de calcular a centralidade de percolação em um problema de estimar probabilidades de intervalos. Para a análise da quantidade de caminhos mínimos necessários, também serão utilizados os resultados da Teoria da Dimensão VC.

O Capítulo 2 apresenta a medida de centralidade de percolação, assim como um algoritmo para calcular essa medida em uma rede. Já nos Capítulos 3 e 4, são introduzidos os conceitos de dimensão VC e ϵ-amostra, além dos resultados necessários para o desenvolvimento deste trabalho. Em seguida, no Capítulo 5, será apresentado o algoritmo proposto, as provas de corretude e um limitante assintótico do tempo de execução. Finalmente, os trabalhos futuros e as conclusões finais serão apresentados no Capítulo 6.

2 Centralidade de Percolação

Este capítulo introduz conceitos de grafos e apresenta uma medida de centralidade que baseia-se na teoria da percolação. Finalmente, na Seção 2.3, é apresentado um algoritmo para o cálculo dessa medida.

2.1 Grafos

Uma rede pode ser representada matematicamente por um grafo. Um *grafo direcionado* G é um par ordenado $(V(G), E(G))$, onde $V(G)$ é um conjunto de vértices e $E(G) \subseteq V(G) \times V(G)$ é um conjunto de arcos. Já um *grafo direcionado com pesos* é um par (G, w), onde G é um grafo direcionado e $w : E(G) \to \mathbb{R}^+$ é uma função que associa um arco e a um número real positivo (peso) $w(e)$. Dado um vértice $v \in V(G)$, o *grau do vértice* v é denotado por $\delta_G(v)$ e representa o número de arcos $e = (u, w) \in E(G)$ tal que $v = u$ ou $v = w$. Logo, as entidades da rede são representadas por vértices no grafo e as interconexões entre as entidades por arcos, que podem ou não possuir pesos.

Dado um grafo direcionado G e um par de vértices $(u, v) \in V(G) \times V(G)$, com $u \neq v$, será chamado de *caminho de u até v* a sequência ordenada de vértices $c_{u,v} = (x_1, \ldots, x_{|c_{u,v}|})$ tal que $x_1 = u$, $x_{|c_{u,v}|} = v$ e $(x_i, x_{i+1}) \in E(G)$, para $1 \leq i < |c_{u,v}|$. O *tamanho do caminho* $c_{u,v}$ será denotado por $|c_{u,v}|$ e o vértice u será chamado de *fonte do caminho* $c_{u,v}$, enquanto que v será chamado de *destino do caminho* $c_{u,v}$. Além disso, todo vértice $w \in c_{u,v}$ tal que $u \neq w \neq v$, será chamado de *vértice interno do caminho* $c_{u,v}$ e $Int(c_{u,v})$ será usado para denotar o conjunto de todos os vértices internos do caminho.

Agora, sendo $c_{u,v}$ um caminho de u até v, o *peso do caminho* $c_{u,v}$ é o somatório dos pesos dos arcos do caminho $\sum_{i=1}^{|c_{u,v}|-1} w(x_i, x_{i+1})$, e que será denotado por $w(c_{u,v})$. Assim, dado um par de vértices $(u, v) \in V(G) \times V(G)$ com $u \neq v$, o *peso do caminho de menor distância* $d_{u,v}$ entre u e v é o peso de um caminho com peso mínimo entre u e v entre todos os caminhos de u até v. Se não existir caminho, então $d_{u,v} = \infty$. Portanto, um caminho $c_{u,v}$ que tem peso $d_{u,v}$ será chamado de *caminho mínimo entre u e v*.

Como podem existir múltiplos caminhos mínimos entre u e v, será denotado por $C_{u,v}$ o conjunto de todos os caminhos $c_{u,v}$ tal que $w(c_{u,v}) = d_{u,v}$. Ainda, a quantidade de caminhos mínimos entre u e v por $\sigma_{u,v} = |C_{u,v}|$. Se não existem caminhos entre u e v, então será definido que $C_{u,v} = \{p_\emptyset\}$, onde $p_\emptyset$ é chamado de *caminho vazio* e tem peso $w(p_\emptyset) = \infty$. Por fim, serão denotados por $\sigma_{u,w}(v)$ a quantidade de caminhos mínimos entre os vértices u e w que têm v como um vértice interno. Isto é, $\sigma_{u,w}(v) = |\{c \mid c \in C_{uw} \text{ e } v \in Int(c)\}|$.

2.2 Centralidade de Percolação

Como visto no ínicio do capítulo, uma parte do estudo de redes, segundo Newman (2010), tem como o objetivo o conceito de *centralidade*. Este conceito, segundo o mesmo autor, quantifica quão importantes são os vértices ou as arestas de uma rede. Medidas de centralidade, como a centralidade de intermediação introduzida por Freeman (1977) e a centralidade de grau, são exemplos de diferentes noções de importância nas redes. Enquanto a primeira considera que entidades importantes são as que estão presentes em mais caminhos mínimos entre pares de entidades na rede, a segunda considera que os vértices mais importantes são os que possuem maior grau.

Esta seção apresenta a centralidade de percolação, introduzida por Piraveenan et al. (2013). Esta medida, a fim de quantificar a importância de um vértice durante a transmissão de uma infestação ou espalhamento de um vírus de computador, baseia-se em estudos como Sander et al. (2002) e Newman (2002), que modelaram infestações de doenças como casos específico da aplicação do processo de percolação em redes.

Segundo o dicionário Priberam (2013), "percolação é a ação ou processo de passar um líquido através de interstícios, para o filtrar ou para com ele extrair componentes solúveis de uma substância.". O estudo desse processo foi introduzido por Broadbent e Hammersley (1957), que atribuíram a aleatoriedade do processo ao meio em que o fluido passa, ou seja, cada "parte do meio" possui um estado (probabilidade) de transmitir ou não o fluido. Por exemplo, na aplicação da percolação em redes, cada parte do meio é um vértice no grafo e cada um desses vértices possui um estado ou probabilidade de ser removido do grafo ou não, pois removendo o vértice do grafo o fluido não será mais transmitido por ele.

Portanto, será denotado por x_i^t o *estado de percolação* do vértice i no tempo t, em que *estado de percolação* é a probabilidade do vértice transmitir um fluido, notícia, doença etc ou ser removido do grafo. Quando o contexto do tempo estiver claro será usado somente x_i. Além disso, o vértice está *totalmente percolado* no tempo t quando $x_i^t = 1$, e *não percolado* quando $x_i^t = 0$. Já quando $0 < x_i^t < 1$, o vértice está *parcialmente percolado.*

Com isso, a centralidade de percolação de um dado vértice v no tempo t é a proporção de caminhos percolados que passam por v, onde um *caminho percolado* é o menor caminho entre um par de vértices, tal que a diferença entre os estados de percolação da fonte e destino é maior que zero. Neste trabalho será utilizado a definição da centralidade de percolação em que os vértices fonte e destino dos caminhos mínimos tem que estar parcialmente percolados. Abaixo a definição matemática da centralidade de percolação é apresentada.

Definição 1. Dado um grafo direcionado com pesos $G = (V(G), E(G))$ e os estados de percolação x_i^t, para todo $i \in V(G)$ no tempo t. A *centralidade de percolação* de um vértice $v \in V(G)$ no instante de tempo t é:

$$pc^t(v) = \sum_{\substack{(u,w)\in V(G)^2 \\ u\neq v\neq w}} \frac{\sigma_{u,w}(v)}{\sigma_{u,w}} \left(\frac{R(x_u^t - x_w^t)}{\sum\limits_{\substack{(f,d)\in V(G)^2 \\ f\neq v\neq d}} R(x_f^t - x_d^t)} \right)$$

onde R é uma função definida como $R(x) = x$ se $x > 0$ e $R(x) = 0$ para $x <= 0$.

Por exemplo, considere a representação do grafo G no tempo $t = 1$ na Figura 2.1, onde $V(G) = \{a, b, c, d, e, f, g, h\}$. Observe que o vértice g é um vértice não percolado, enquanto que a possui $x_a^1 = 0.1$, logo é um vértice parcialmente percolado.

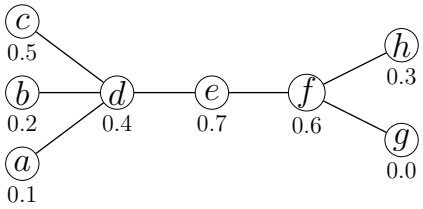


Figura 2.1: Grafo de exemplo

v	a	b	c	d	e	f	g	h
$pc^1(v)$	0	0	0	0.64	0.53	0.51	0	0

Tabela 2.1: Centralidade de Percolação para o grafo da Figura 2.1

Aplicando a Definição 1 em cada vértice do grafo, é obtido como resultado a Tabela 2.1, indicando que o vértice mais central do grafo seria o vértice d, pois possui maior centralidade de percolação. A Figura 2.2 mostra todos os caminhos mínimos entre pares de vértices no qual o vértice d é vértice interno do caminho.

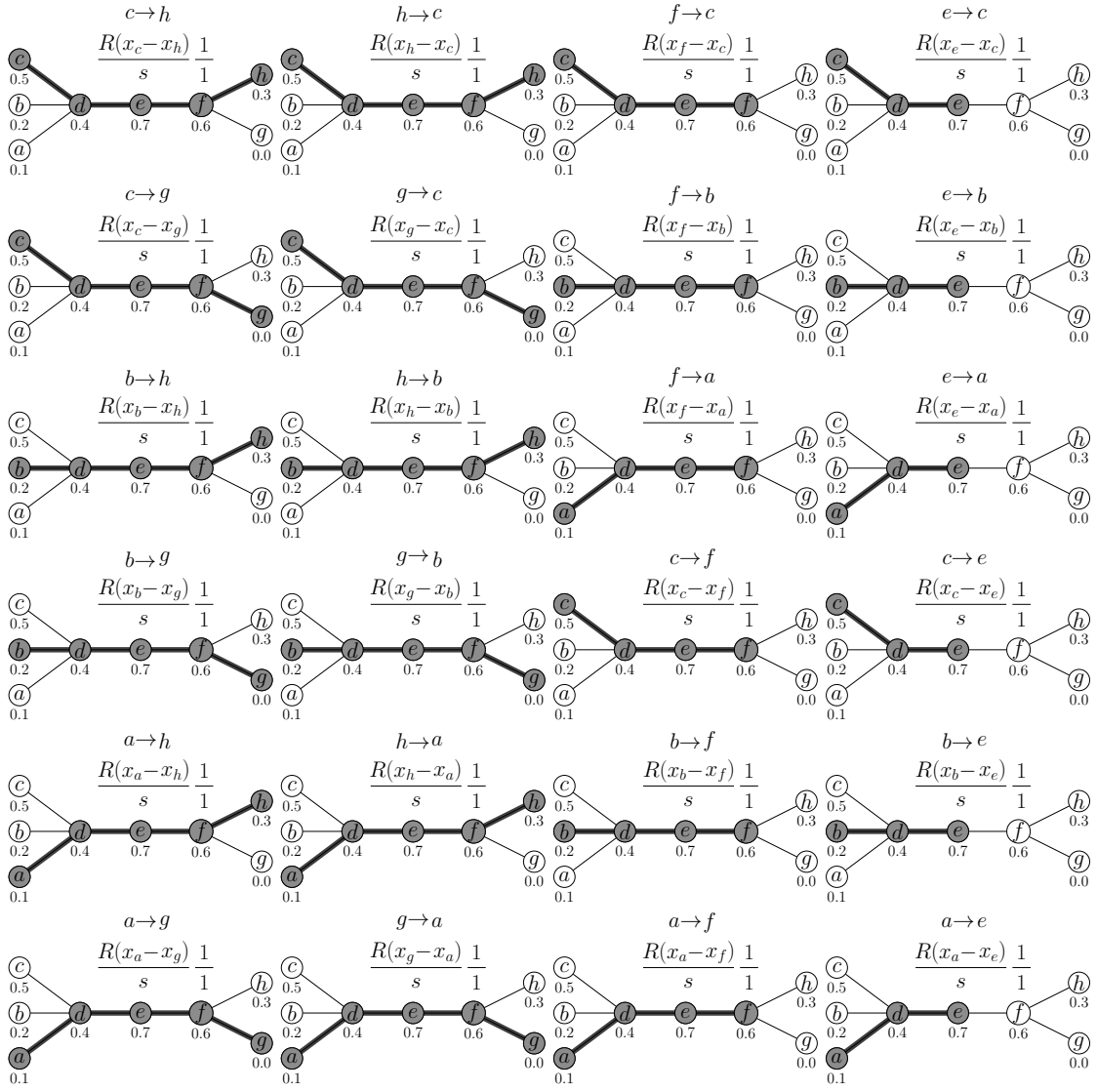


Figura 2.2: Todos os caminhos mínimos entre pares de vértices em que o vértice d atua como vértice interno. Logo acima de cada caminho é descrito o valor que é adicionado a $pc^1(d)$, onde $s = \sum_{u \neq d \neq w} R(x_u - x_w)$. Veja que $\frac{\sigma_{u,w}(v)}{\sigma_{u,w}}$ é sempre igual a 1 neste grafo, pois $|C_{u,v}| = 1$ para todos os pares de vértices. Além disso, observe que todo caminho que tem como fonte o vértice g e como destino qualquer outro vértice v é desconsiderado, pois como $x_g = 0$ então $R(x_g - x_v) = 0$.

Portanto, supondo que este grafo representa a infestação de uma doença em um conjunto de cidades, a cidade d deveria ter uma atenção especial no tempo $t = 1$ para conter a infestação, de acordo com esta medida de centralidade.

2.3 Algoritmo Conhecido

Esta seção apresenta um algoritmo para calcular a centralidade de percolação da Definição 1. Segundo Piraveenan et al. (2013), dado um grafo $G = (V(G), E(G))$, e os estados de percolação x_i^t para todo $i \in V(G)$ no instante de tempo t, o algoritmo para computar a centralidade de percolação tem complexidade de tempo $\Theta(n^3)$, onde $n = |V(G)|$ é o número de vértices do grafo. Este algoritmo é uma modificação do algoritmo que calcula a centralidade de intermediação e é apresentado no Algoritmo 1. Neste algoritmo, a função *floyd_warshall* é uma chamada para o algoritmo de Floyd–Warshall (Cormen et al. (2009)) e retorna duas matrizes, uma matriz $dist$ contendo a menor distância entre os vértices e a matriz σ contendo o número de caminhos mínimos entre os vértices.

Algoritmo 1: CentralidadeDePercolacao(G,x)

Dados: Grafo $G = (V(G), E(G))$ com $|V(G)| = n$ e os estados de percolação x_i^t, para $1 \leq i \leq n$.

Resultado: A centralidade de percolação dos vértices $v \in V(G)$.

```
1  dist,σ ← floyd_warshall (G)
2  para v ∈ V faça
3      soma ← 0;
4      para u ∈ V faça
5          para w ∈ V faça
6              se u ≠ w ≠ v então
7                  se dist[u][w] = dist[u][v] + dist[v][w] então
8                      σ_uwv ← σ[u][v]σ[v][w];
9                  fim
10                 dif ← x[u] − x[w];
11                 se dif > 0 então
12                     pc(v) ← pc(v) + dif σ_uwv / σ[u][w];
13                     soma ← soma + dif;
14                 fim
15             fim
16         fim
17     fim
18     pc(v) ← pc(v) / soma;
19 fim
20 retorna pc ;
```

Portanto, utilizando essas matrizes, é calculado na linha 8 o número de caminhos mínimos entre u e w que passam por v. Em seguida, esse resultado é utilizado para calcular $\frac{\sigma_{u,w}(v)}{\sigma_{u,w}} R(x_u - x_w)$.

Brandes (2001) apresenta um algoritmo de complexidade de tempo $O(nm)$, onde $n = |V(G)|$ e $m = |E(G)|$, para o cálculo da centralidade de intermediação. Porém, segundo Piraveenan et al. (2013), não é possível modificar esse algoritmo para o cálculo da centralidade de percolação que considera que a fonte e o destino dos caminhos estejam percolados totalmente

ou parcialmente. Essa impossibilidade ocorre devido ao algoritmo de Brandes não manter a referência do destino quando está calculando o menor caminho entre pares de vértices e, portanto, não seria possível calcular $R(x_u - x_w)$.

Este capítulo introduziu conceitos da teoria dos grafos, que serão utilizados posteriormente neste trabalho. Além disso, a centralidade de percolação e um algoritmo que realiza o cálculo desta medida foram apresentados.

3 Dimensão Vapnik–Chervonenkis

A Lei dos Grandes Números diz que a frequência de um evento em uma sequência de observações converge para a probabilidade do evento. Entretanto, em muitas aplicações, é preciso que exista essa convergência quando mais de um evento simultaneamente está sendo considerado. A Dimensão Vapnik–Chervonenkis (VC) é uma medida de complexidade de um conjunto de eventos e foi apresentada em Vapnik e Chervonenkis (1971). Além disso, nesse mesmo trabalho, foi demonstrado que a Dimensão VC é condição suficiente para a convergência das frequências relativas de um conjunto de eventos, independentemente da distribuição.

Esta medida possui aplicações na área de Geometria Computacional como visto em Matoušek (2002). Além disso, ela foi introduzida na área de Teoria do Aprendizado Computacional no trabalho de Haussler e Welzl (1986) e Blumer et al. (1989) e tornou-se um elemento central no modelo *Probably Approximately Correct* (PAC), um modelo matemático para aprendizagem apresentado por Valiant (1984).

Este capítulo apresenta o conceito de espaço de intervalos, projeção, despedaçar de subconjuntos e Dimensão VC. Por fim, é demonstrado o Teorema Sauer–Shelah, que será utilizado no Capítulo 4.

3.1 Dimensão VC

A Dimensão VC é definida sobre um espaço de intervalos.

Definição 2. Um *espaço de intervalos* é um par $R = (X, \mathcal{I})$ onde:

- X é um conjunto (finito ou infinito) de pontos.
- $\mathcal{I}$ é uma família de subconjuntos de X, chamados de intervalos.

Por exemplo, seja $X = \{1, 2, 3\}$ e $\mathcal{I} = \{\{1\}, \{1, 2\}, \{1, 2, 3\}, \{2\}, \{2, 3\}, \{3\}\}$, então $R = (X, \mathcal{I})$ é um espaço de intervalos sobre o conjunto dos números de 1 a 3. Observe agora, que dado um subconjunto $S \subseteq X$, pode-se fazer a interseção desse conjunto com todo intervalo $I \in \mathcal{I}$. Obtendo com isso, um conjunto de subconjuntos de S, que é chamado de *projeção* de $\mathcal{I}$ em S.

No exemplo, se $S = \{1, 3\}$, então os resultados das interseções são:

$$\{1, 3\} \cap \{1\} = \{1\}$$

$$\{1, 3\} \cap \{1, 2\} = \{1\}$$

$$\{1, 3\} \cap \{1, 2, 3\} = \{1, 3\}$$

$$\{1, 3\} \cap \{2\} = \emptyset$$

$$\{1, 3\} \cap \{2, 3\} = \{3\}$$

$$\{1,3\} \cap \{3\} = \{3\}$$

Logo, $\mathcal{I}_{\{1,3\}} = \{\emptyset, \{1\}, \{1,3\}, \{3\}\}$. Note que, neste caso, $\mathcal{I}_{\{1,3\}} = 2^S$, onde 2^S denota todos os subconjuntos possíveis de S. Agora, por exemplo, considerando $S = \{1,2,3\}$, é observado que $\mathcal{I}_{\{1,2,3\}}$ não é igual a $2^{\{1,2,3\}}$, pois não é possível gerar o subconjunto $\{1,3\}$ fazendo somente as interseções com os intervalos. Com base nisso, quando a projeção de $\mathcal{I}$ em S é igual a todos os subconjuntos de S, este subconjunto S é dito ser *despedaçado* por I.

Definição 3. Seja $R = (X, \mathcal{I})$ um espaço de intervalos. Um conjunto $S \subseteq X$ é *despedaçado* por $\mathcal{I}$ se $|\mathcal{I}_S| = 2^{|S|}$.

Dado esses conceitos, a Dimensão VC de um espaço de intervalos é definida abaixo.

Definição 4. A *Dimensão Vapnik-Chervonenkis (VC)* de um espaço de intervalos $R = (X, \mathcal{I})$, denotada por $VCDim(R)$, é a maior cardinalidade de um conjunto $S \subseteq X$ que é despedaçado por $\mathcal{I}$.

$$VCDim(R) = max\{d : \exists |S| = d \text{ e } |\mathcal{I}_S| = 2^d\}$$

Se existirem conjuntos finitos arbitrariamente grandes que são despedaçados por $\mathcal{I}$, então $VCDim(R) = \infty$.

Portanto, voltando ao exemplo apresentado, pode-se observar que $VCDim(R) = 2$, pois o único subconjunto $S = \{1,2,3\}$ de tamanho 3 não é despedaçado por $\mathcal{I}$ e foi mostrado que existe um subconjunto de tamanho 2 que é despedaçado por $\mathcal{I}$. Observe que para mostrar que a Dimensão VC de um espaço de intervalos é d, é necessário mostrar que existe um subconjunto S de X tal que $|S| = d$ e que é despedaçado por $\mathcal{I}$ e, além disso, que todo subconjunto de tamanho $d + 1$ não é despedaçado $\mathcal{I}$.

Se o número de intervalos em um espaço de intervalos for finito, é possível limitar superiormente a Dimensão VC deste espaço de intervalos por $\log_2(|\mathcal{I}|)$, como mostra a Observação abaixo.

Observação 1. *Seja $R = (X, \mathcal{I})$ um espaço de intervalos tal que $|\mathcal{I}| < \infty$. Neste caso temos que*

$$VCDim(R) \leq \log_2(|\mathcal{I}|)$$

Demonstração. Considere que $VCDim(R) = d$, então existe um subconjunto $S \subseteq X$ de tamanho d tal que $|\mathcal{I}_S| = 2^{|S|} = 2^d$. Portanto, existem pelo menos 2^d intervalos no conjunto $\mathcal{I}$. Assim, $|\mathcal{I}| \geq 2^d$ e daí $\log_2(|\mathcal{I}|) \geq d$, ou seja, $VCDim(R) \leq \log_2(|\mathcal{I}|)$. □

3.2 Exemplos

3.2.1 Intervalos fechados em $\mathbb{R}$

Um intervalo fechado em $\mathbb{R}$ é um conjunto $[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$. Será definido por $R = (\mathbb{R}, \mathcal{I})$ o espaço de intervalos fechados em $\mathbb{R}$, no qual $\mathcal{I} = \{[a, b] \mid a, b \in \mathbb{R}\}$. Em seguida, será mostrado que $VCDim(R) = 2$.

Para chegar a essa conclusão, é necessário mostrar que algum conjunto de 2 elementos é despedaçado por $\mathcal{I}$ e que nenhum conjunto de 3 ou mais elementos é despedaçado por $\mathcal{I}$. Observe que o conjunto $S = \{2,4\}$ pode ser despedaçado por $\mathcal{I}$, pois a interseção com o intervalo $[0,1]$ resulta no conjunto vazio; a interseção de S com o intervalo $[1,3]$ é $\{2\}$; a interseção de S com o intervalo $[3,5]$ é $\{4\}$;e a interseção de de S com o intervalo $[1,5]$ é $\{2,4\}$. Assim, o

tamanho da projeção de $\mathcal{I}$ em $\{2,4\}$ é $|\mathcal{I}_{\{2,4\}}| = 2^{|\{2,4\}|} = 4$; portanto, S é despedaçado por $\mathcal{I}$, e a Dimensão VC de R é pelo menos 2.

Agora, será visto que a Dimensão VC de R não pode ser 3 ou mais. Considere um conjunto $S = \{a, b, c\}$ de três elementos e, sem perda de generalidade, considere que $a < b < c$. Observe então, que é possível gerar 7 dos 8 conjuntos necessários para despedaçar S usando $\mathcal{I}$. Porém, não é possível fazer com que o resultado da interseção de S com algum intervalo em $\mathcal{I}$ seja igual a $\{a, c\}$, pois todo intervalo que possui $\{a, c\}$ também possui b. Esse argumento vale também para conjuntos com mais de três elementos.

Na Figura 3.1, são ilustrados alguns intervalos fechados nos reais no intervalo de 0 a 4. A reta real (a) da figura possui intervalos fechados cuja a diferença entre a e b é igual a 0, já a reta real (b) representa os intervalos com diferença igual a 1 e assim por diante até a reta (d). Pode-se observar na reta (c) que não é possível escolher um intervalo que não possua o número 2.

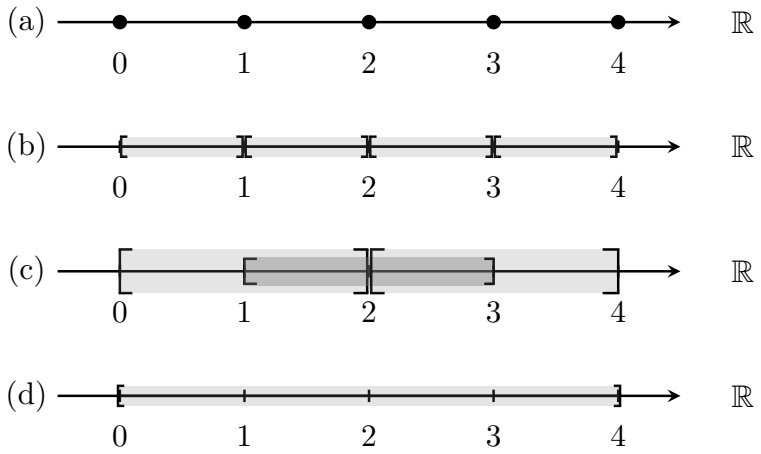


Figura 3.1: Intervalos fechados da reta dos reais.

3.2.2 Conjuntos Convexos

Um conjunto C é convexo se $\theta x_1 + (1 - \theta)x_2 \in C$ para todo $x_1, x_2 \in C$ e para todo $\theta \in [0, 1]$. Seja $X = \mathbb{R}^2$ e $\mathcal{I} = \{C \mid C \subseteq \mathbb{R}^2 \text{ e } C \text{ é convexo}\}$, então $R = (\mathbb{R}^2, \mathcal{I})$ será definido como o espaço de intervalos de todos os conjuntos convexos no plano. Feito isso, será demonstrado que $VCDim(R) = \infty$. Considere $S_n = \{x_1, \ldots, x_n\}$ o conjunto de n pontos co-circulares. Qualquer subconjunto $Y \subseteq S_n$ com $Y \neq \emptyset$ define um conjunto convexo que não inclui nenhum ponto em $S_n \backslash Y$ e portanto Y é incluído na projeção de $\mathcal{I}$ em S_n. É fácil ver também que o conjunto vazio está nessa projeção, basta pegar um ponto que não está em S_n. Portanto, para qualquer número de pontos n, o conjunto S_n é quebrado e a Dimensão VC é, portanto, infinita.

A Figura 3.2 ilustra um exemplo quando $n = 4$ e $S = \{x_1, x_2, x_3, x_4\}$. Em cada circuferência da figura, é apresentado o conjunto $I \in \mathcal{I}$, que será utilizado na projeção de $\mathcal{I}$ em S que despedaça S.

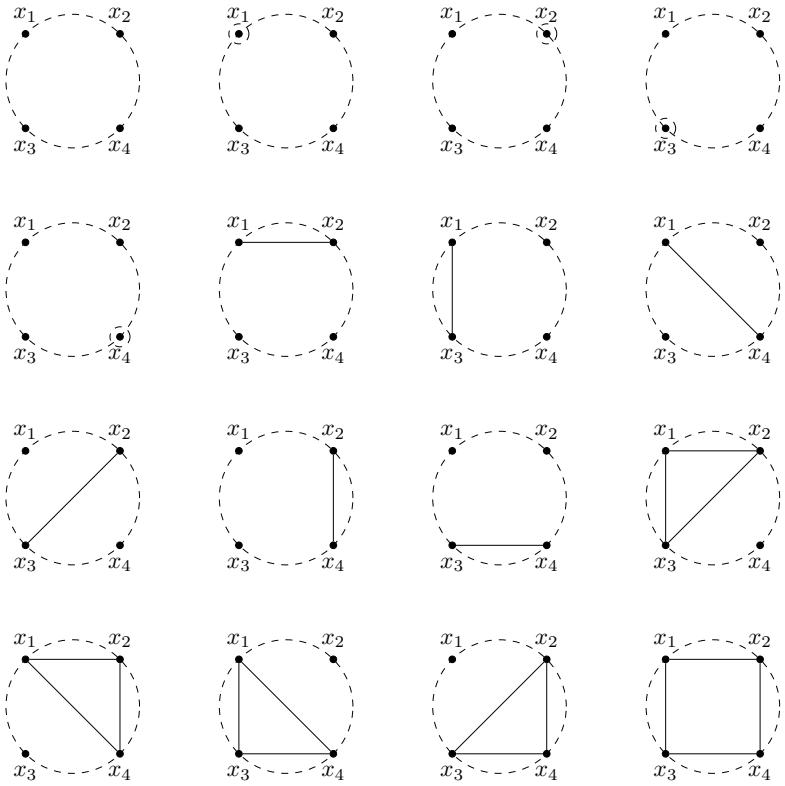


Figura 3.2: Conjunto de 4 pontos quebrados por $\mathcal{I}$

3.3 Função $\prod_{\mathcal{I}}(m)$

Esta seção define a função $\prod_{\mathcal{I}}(m)$ para um espaço de intervalos, que para Kearns et al. (1994) pode ser vista como uma medida de complexidade do espaço. Após isso, é demonstrado que essa função é limitada por um polinômio, cujo grau é a Dimensão VC do espaço de intervalos.

Portanto, dado um espaço de intervalos $R = (X, \mathcal{I})$ a função $\prod_{\mathcal{I}}(m)$ será definida como:

Definição 5. Para qualquer número $m \in \mathbb{N}$:

$$\prod\nolimits_{\mathcal{I}}(m) = max\{|\mathcal{I}_S| : |S| = m\}$$

Primeiramente, será provado no Teorema 1 que $\prod_{\mathcal{I}}(m)$ é limitada superiormente pela função $\phi_d(m)$, definida abaixo. Na sequência, será mostrado no Lema 1 que $\phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$ e, no Teorema 2, que o número de intervalos de um espaço de intervalos $R = (X, \mathcal{I})$ com $|X| = n$ e $VCDim(R) = d$ é limitado superiormente por n^d.

Definição 6. Para quaisquer naturais m e d, a função $\phi_d(m)$ é definida recursivamente por:

$$\phi_d(m) = \begin{cases} \phi_d(m-1) + \phi_{d-1}(m-1), & \text{se } d \neq 0 \text{ e } m \neq 0; \\ 1 & \text{se } d = 0 \text{ ou } m = 0. \end{cases} \tag{3.1}$$

Teorema 1 (Kearns et al.)**.** *Seja $R = (X, \mathcal{I})$ um espaço de intervalos e $VCDim(R) = d$. Então, para qualquer $m \in \mathbb{N}$:*

$$\prod\nolimits_{\mathcal{I}}(m) \leq \phi_d(m)$$

Demonstração. Será provado que $\prod_{\mathcal{I}}(m) \leq \phi_d(m)$ por indução em d e em m. Portanto, como base da indução: (1) d é arbitrário e $m = 0$ e (2) $d = 0$ e m é arbitrário. Para o caso (1), o único $S \subseteq X$ tal que $|S| = 0$ é o conjunto vazio e como $\mathcal{I}_\emptyset = \{\emptyset\}$, então $\prod_{\mathcal{I}}(m) = 1$. Por outro lado, $\phi_d(0) = 1$, logo $\prod_{\mathcal{I}}(0) = \phi_d(0) = 1$. Já para o caso (2), como $d = 0$, então $|\mathcal{I}| = 1$ e daí $\prod_{\mathcal{I}}(m) = 1 = \phi_0(m)$ para todo $m \in \mathbb{N}$.

Agora, como hipótese de indução, será assumido que para $m' \leq m$ e $d' \leq d$ com pelo menos uma das desigualdades estritas, vale a desigualdade $\prod_{\mathcal{I}}(m') \leq \phi_{d'}(m')$. Na sequência, será mostrado que usando essa hipótese de indução é possível mostrar que a desigualdade vale para d e m.

Dado qualquer conjunto $S \subseteq X$ tal que $|S| = m$, seja $x \in S$ um ponto qualquer. Primeiramente, será calculado $|\mathcal{I}_{S-\{x\}}|$, o que é fácil, pois como $|S - \{x\}| = m - 1$ e daí pela hipótese de indução é obtido que $|\mathcal{I}_{S-\{x\}}| \leq \phi_d(m-1)$. Observe que a diferença entre $\mathcal{I}_S$ e $\mathcal{I}_{S-\{x\}}$ é que os conjuntos em $\mathcal{I}_S$ que eram diferenciados somente pelo elemento x se tornam somente um conjunto único em $\mathcal{I}_{S-\{x\}}$. Com base nisso, será definido um novo espaço de intervalos $R' = (X - \{x\}, \mathcal{I}')$ onde:

$$\mathcal{I}' = \{I \in \mathcal{I}_S \mid x \notin I \text{ e } I \cup \{x\} \in \mathcal{I}_S\}$$

Observe que $|\mathcal{I}'|$ conta o número de conjuntos que se transformam em um só quando é calculado $\mathcal{I}_{S-\{x\}}$. Observe também que $\mathcal{I}' = \mathcal{I}'_{S-\{x\}}$, pois $\mathcal{I}'$ consiste somente de subconjuntos de $S - \{x\}$. Assim,

$$|\mathcal{I}_S| = |\mathcal{I}_{S-\{x\}}| + |\mathcal{I}'_{S-\{x\}}|$$

Agora será mostrado que $VCDim(R') = d - 1$. Para isso, suponha por contradição que $\mathcal{I}'$ despedaça um conjunto S de tamanho d em $X - \{x\}$. Considere agora o conjunto $S \cup \{x\}$ em $\mathcal{I}$. Para qualquer $I \in \mathcal{I}'$, ambos I e $I \cup \{x\}$ estão em $\mathcal{I}$, e assim pode-se obter ambos os conjuntos $(S \cup \{x\}) \cap I = S \cap I$ e $(S \cup \{x\}) \cap (I \cup \{x\}) = S \cup \{x\}$ na projeção de $\mathcal{I}$ em S. Mas então $\mathcal{I}$ tem que despedaçar o conjunto $S \cup \{x\}$, que tem tamanho $|S \cup \{x\}| = |S| + |\{x\}| = d + 1$, mas isso contradiz a suposição do teorema que diz que $VCDim(R) = d$. Portanto, aplicando a hipótese de indução:

$$|\mathcal{I}'| = |\mathcal{I}'_{S-\{x\}}| \leq \phi_{d-1}(m-1)$$

Finalmente, pela Definição 6:

$$\phi_d(m-1) + \phi_{d-1}(m-1) = \phi_d(m)$$

□

Lema 1 (Kearns et al.)**.** $\phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$

Demonstração. Será provado por indução em d e m. Para a base da indução, observe que quando $d = 0$ e m é arbitrário:

$$\phi_0(m) = 1 = \sum_{i=0}^{0} \binom{m}{i} = \binom{m}{0}$$

Por outro lado quando $m = 0$ e d é arbitrário:

$$\phi_d(0) = 1 = \sum_{i=0}^{d} \binom{0}{i}$$

Agora, como hipótese de indução, será assumido que

$$\phi_d(m-1) = \sum_{i=0}^{d} \binom{m-1}{i}$$

e

$$\phi_{d-1}(m-1) = \sum_{i=0}^{d-1} \binom{m-1}{i}$$

E daí, pela Definição 6:

$$\phi_d(m) = \phi_d(m-1) + \phi_{d-1}(m-1)$$

Finalmente, pela hipótese de indução:

$$\begin{aligned}
\phi_d(m-1) + \phi_{d-1}(m-1) &= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{j=1}^{d} \binom{m-1}{j-1} \\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{j=0}^{d} \binom{m-1}{j-1} \\
&= \sum_{i=0}^{d} \left(\binom{m-1}{i} + \binom{m-1}{i-1} \right) \\
&= \sum_{i=0}^{d} \binom{m}{i}
\end{aligned}$$

□

Teorema 2 (Sauer–Shelah (Sauer, 1972; Shelah, 1972))**.** *Seja $R = (X, \mathcal{I})$ um espaço de intervalos com $|X| = n$ e $VCDim(R) = d$. Então $|\mathcal{I}| \leq n^d$.*

Demonstração. Como o conjunto X tem tamanho finito, então o maior e único subconjunto $S \subseteq X$ de tamanho n é o próprio conjunto X. Além disso, observe que a maior projeção de $\mathcal{I}$ em um subconjunto de tamanho n é igual ao próprio $\mathcal{I}$, logo $\mathcal{I} = \prod_{\mathcal{I}}(n) = \mathcal{I}_X$. Portanto, pelo Teorema 1:

$$|\mathcal{I}| = \prod\nolimits_{\mathcal{I}}(n) = |\mathcal{I}_X| \leq \phi_d(n)$$

E daí, usando o Lema 1:

$$
\begin{aligned}
\phi_d(n) &\leq \sum_{i=0}^{d} \binom{n}{i} \\
&= \sum_{i=0}^{d} \frac{n!}{(n-i)!i!} = \sum_{i=0}^{d} \frac{n(n-1)\dots(n-i+1)}{i!} \\
&\leq \sum_{i=0}^{d} \frac{n(n)\dots(n)}{i!} = \sum_{i=0}^{d} \frac{n^{i-1}}{i!} \\
&\leq \sum_{i=0}^{d} \frac{n^i}{i!} \leq \sum_{i=0}^{d} \frac{n^d}{d!} = d\frac{n^d}{d!} = \frac{n^d}{(d-1)!} \leq n^d
\end{aligned}
$$

□

4 ϵ-amostra

O levantamento por amostragem, segundo Silva (2015), permite a obtenção de informações a respeito de valores desconhecidos da população, por meio de amostras (partes da população). Média, variância e tamanho de subconjuntos são exemplos de valores populacionais desconhecidos que se pode obter. Porém, esse tipo de levantamento pode conter erros (Bolfarine e Bussab, 2005), como o erro amostral, que corresponde à diferença entre o valor desconhecido da população e o valor calculado utilizando a amostra. Note que esse erro varia dependendo da amostra escolhida e do valor desconhecido que está sendo calculado.

Este capítulo terá como objeto de estudo amostras que possuem erro amostral baixo na estimativa de tamanho de subconjuntos, as chamadas ϵ-amostra. Além disso, a Seção 4.2 relaciona Dimensão VC com ϵ-amostra, limitando inferiormente o tamanho da amostra para que seja uma ϵ-amostra com alta probabilidade.

4.1 Definição

Dado um espaço de intervalos $R = (X, \mathcal{I})$, será chamado de população o conjunto X e de subconjunto da população os intervalos $\mathcal{I}$. Além disso, seja $\mathcal{D}$ uma distribuição de probabilidade sobre a população X e $\epsilon \in \mathbb{R}$ um erro tolerável. Será denotado por $\Pr_{\mathcal{D}}(I)$ a probabilidade do subconjunto I de acordo com a distribuição $\mathcal{D}$. Com base nisso, Mitzenmacher e Upfal (2017) definem ϵ-amostra como:

Definição 7. Seja $R = (X, \mathcal{I})$ um espaço de intervalos, e seja $\mathcal{D}$ uma distribuição de probabilidade em X. Um conjunto $S \subseteq X$ é uma *ϵ-amostra* para X com respeito a $\mathcal{D}$ se para todos os conjuntos $I \in \mathcal{I}$,

$$\left| \Pr_{\mathcal{D}}(I) - \frac{|S \cap I|}{|S|} \right| \leq \epsilon$$

Como exemplo, suponha que um país vai eleger um novo presidente e um instituto de pesquisa deseja fazer uma estimativa do número de eleitores de cada candidato com baixa porcentagem de erro. Neste país, estão concorrendo para o cargo os candidatos C_1, C_2, C_3 e C_4. Seja P o conjunto de todas as pessoas deste país e P_{C_1} o subconjunto de pessoas que votam no candidato C_1, P_{C_2} o subconjunto de pessoas que votam no candidato C_2 e assim por diante até C_4. O instituto então deseja estimar $|P_{C_i}|$, para $1 \leq i \leq 4$, utilizando somente uma amostra das pessoas do país.

O instituto então amostrou uma quantidade m de pessoas (que já tinham o candidato em que iriam votar decidido) e contou, para cada candidato, o número de eleitores. Após isso, foi calculado a porcentagem de pessoas da amostra que votariam em cada candidato. Logo após as eleições nesse país, o instituto observou que o erro da sua pesquisa foi de 1% e concluiu que a pesquisa foi um sucesso.

Portanto, dado o conjunto $\mathcal{I} = \{P_{C_i} \mid 1 \leq i \leq 4\}$ e definindo $E = (P, \mathcal{I})$ como espaço de intervalos da eleição. Considere então, que $\mathcal{D}$ é a distribuição dos votos sobre a população P. Logo, o conjunto de pessoas amostradas pelo instituto é uma 0.01-amostra para o espaço de intervalos $E = (P, \mathcal{I})$ e a distribuição de votos $\mathcal{D}$. Portanto, o objetivo inicial do instituto era obter uma ϵ-amostra para esse espaço de intervalos e distribuição.

4.2 Teorema ϵ-amostra

No exemplo da seção anterior, o instituto conseguiu uma ϵ-amostra de tamanho m para o espaço de intervalos $E = (P, \mathcal{I})$ e distribuição $\mathcal{D}$. Nesta seção, será apresentado um limite no tamanho da amostra m para que se tenha uma ϵ-amostra com alta confiabilidade. Para obter esse limite, serão utilizados os conceitos de Dimensão VC, assim como o Teorema Sauer–Shelah.

Teorema 3 (Mitzenmacher e Upfal)**.** *Seja $R = (X, \mathcal{I})$ um espaço de intervalos tal que $VCDim(R) = d$ e seja $\mathcal{D}$ uma distribuição de probabilidade em X. Para qualquer $0 < \epsilon < \frac{1}{2}$ e $0 < \delta < \frac{1}{2}$, existe*

$$m \geq \frac{32d}{\epsilon^2} ln \frac{64d}{\epsilon} + \frac{16}{\epsilon^2} ln \frac{2}{\delta}$$

tal que uma amostra aleatória com respeito a $\mathcal{D}$ de tamanho maior ou igual a m é uma ϵ-amostra para X com probabilidade $1 - \delta$.

Demonstração. Considere uma amostra M amostrada de acordo com a distribuição $\mathcal{D}$, tal que $|M| = m$. Seja agora o evento E_1 o evento da amostra não ser uma ϵ-amostra para o espaço de intervalos $R = (X, \mathcal{I})$. Portanto, pela Definição 7:

$$E_1 = \left\{ \exists I \in \mathcal{I} \mid \left| Pr_{\mathcal{D}}(I) - \frac{|M \cap I|}{|M|} \right| > \epsilon \right\}$$

Será mostrado neste teorema que $\Pr(E_1) \leq \delta$. Primeiramente, uma outra amostra T de tamanho m amostrada de acordo com à distribuição $\mathcal{D}$ será construída. Agora, considere o seguinte evento E_2, onde existe pelo menos um $I \in \mathcal{I}$ tal que I não é bem aproximado por M, mas é bem aproximado por T.

$$E_2 = \left\{ \exists I \in \mathcal{I} \mid \left| \Pr_{\mathcal{D}}(I) - \frac{|M \cap I|}{|M|} \right| > \epsilon \text{ e } \left| \Pr_{\mathcal{D}}(I) - \frac{|I \cap T|}{|T|} \right| \leq \frac{\epsilon}{2} \right\}$$

Com isso, o lema abaixo limita $\Pr(E_1)$ utilizando $\Pr(E_2)$.

Lema 2 (Mitzenmacher e Upfal)**.**

$$\Pr(E_2) \leq \Pr(E_1) \leq 2 \Pr(E_2)$$

Demonstração. Para mostrar que $\Pr(E_2) \leq \Pr(E_1) \leq 2 \Pr(E_2)$ é preciso provar que $\Pr(E_2) \leq \Pr(E_1)$ e $\Pr(E_1) \leq 2 \Pr(E_2)$. No primeiro caso é fácil ver que $E_2 \subseteq E_1$, portanto $\Pr(E_2) \leq \Pr(E_1)$. Já para o segundo caso, basta mostrar que $\frac{\Pr(E_2)}{\Pr(E_1)} \geq \frac{1}{2}$. Mas observe que:

$$\frac{\Pr(E_2)}{\Pr(E_1)} = \frac{\Pr(E_2 \cap E_1)}{\Pr(E_1)} = \Pr(E_2 \mid E_1)$$

Mas se E_1 acontece, então existe um intervalo I' tal que $\left| \Pr_{\mathcal{D}}(I') - \frac{|M \cap I'|}{|M|} \right| > \epsilon$. Logo, pode-se definir o evento E_1', em que só existe o intervalo I':

$$E_1' = \left\{ \left| \Pr_{\mathcal{D}}(I') - \frac{|M \cap I'|}{|M|} \right| > \epsilon \right\}$$

Portanto, como podem existir mais que um intervalo em E_1 que possuem erro maior que ϵ, então:

$$\begin{aligned} \Pr(E_2 \mid E_1) \geq \Pr(E_2 \mid E_1') &= \Pr\left(\left| \Pr_{\mathcal{D}}(I') - \frac{|T \cap I'|}{|T|} \right| \leq \frac{\epsilon}{2} \right) \\ &= \Pr\left(\left| m \Pr_{\mathcal{D}}(I') - |T \cap I'| \right| \leq \frac{m\epsilon}{2} \right) \end{aligned}$$

Note que para um intervalo fixo I' e para uma amostra T, a variável aleatória $|T \cap I'|$ segue uma distribuição binomial $B(m, \Pr_{\mathcal{D}}(I'))$ (Morettin e Bussab (2013)) tal que $E[|T \cap I'|] = m \Pr_{\mathcal{D}}(I') \leq m$ e aplicando o Limitante de Chernoff (Mitzenmacher e Upfal (2017)):

$$\begin{aligned} \Pr\left(||T \cap I'| - m \Pr_{\mathcal{D}}(I')| \leq \frac{m\epsilon}{2} \right) &= 1 - \Pr\left(||T \cap I'| - m \Pr_{\mathcal{D}}(I')| > \frac{m\epsilon}{2} \right) \\ &\geq 1 - 2e^{-m \Pr_{\mathcal{D}}(I')(\frac{\epsilon}{2})^2/3} \\ &= 1 - 2e^{-m \Pr_{\mathcal{D}}(I')\epsilon^2/12} \\ &\geq 1 - 2e^{-m\epsilon^2/12} \\ &\geq 1 - \frac{1}{2} = \frac{1}{2} \end{aligned}$$

para $m \leq \frac{24}{\epsilon}$. Finalmente:

$$\frac{\Pr(E_2)}{\Pr(E_1)} = \Pr(E_2 \mid E_1) \geq \Pr\left(\left| m \Pr_{\mathcal{D}}(I') - m|T \cap I'| \right| \leq \frac{m\epsilon}{2} \right) \geq \frac{1}{2}$$

□

Com base nisso, o evento E_2 será limitado por outro evento ainda maior E_2':

$$E_2' = \left\{ \exists I \in \mathcal{I} \mid ||T \cap I| - |M \cap I|| \geq \frac{\epsilon}{2} m \right\}$$

Lema 3 (Mitzenmacher e Upfal)**.**

$$E_2 \subseteq E_2'$$

Demonstração. Assumindo que um conjunto I satisfaz as condições de E_2, ou seja,

$$\left| |I \cap M| - m \Pr_{\mathcal{D}}(I) \right| \geq \epsilon m \tag{4.1}$$

e

$$\left| |I \cap T| - m \Pr_{\mathcal{D}}(I) \right| \leq \frac{\epsilon m}{2} \tag{4.2}$$

Portanto, subtraindo 4.1 de 4.2:

$$\left| |I \cap M| - m \Pr_{\mathcal{D}}(I) \right| - \left| |I \cap T| - m \Pr_{\mathcal{D}}(I) \right| \geq \frac{\epsilon m}{2}$$

e usando a desigualdade triangular[1]:

[1] $|x - y| \geq ||x| - |y||$

$$
\begin{aligned}
||I \cap M| - |I \cap T|| &= \left||I \cap M| - m \Pr_{\mathcal{D}}(I) - |I \cap T| + m \Pr_{\mathcal{D}}(I)\right| \\
&\geq \left||I \cap M| - m \Pr_{\mathcal{D}}(I)\right| - \left||I \cap T| - m \Pr_{\mathcal{D}}(I)\right| \\
&\geq \frac{\epsilon m}{2}
\end{aligned}
$$

□

Observe que o evento E_2' só depende dos elementos de $M \cup T$. Portanto, esse fato será utilizado para limitar E_2'.

Lema 4 (Mitzenmacher e Upfal)**.**

$$\Pr(E_2) \leq \Pr(E_2^{'}) \leq (2m)^d e^{-\epsilon^2 m/8}$$

Demonstração. É necessário mostrar $\Pr(E_2) \leq Pr(E_2')$ e $\Pr(E_2') \leq (2m)^d e^{-\epsilon^2 m/8}$. No primeiro caso, como $E_2 \subseteq E_2'$ pelo Lema 3, então $\Pr(E_2) \leq \Pr(E_2')$. Já para o segundo caso, como M e T são amostras aleatórias, é possível assumir que serão amostrados $2m$ elementos $Z = z_1, \ldots, z_{2m}$ e então particionados aleatoriamente em dois conjuntos de tamanho m.

Como Z já é uma amostra aleatória, qualquer partição que é independente dos valores dos elementos vai gerar duas amostras aleatórias. A seguinte partição será utilizada: para cada par de amostras z_{2i-1} e z_{2i} para $i = 1, \ldots, m$ com probabilidade $\frac{1}{2}$ será adicionado z_{2i-1} em T e z_{2i} em M, caso contrário será adicionado z_{2i-1} em M e z_{2i} em T.

Agora considere um intervalo fixo $I' \in \mathcal{I}$, com base nisso será definido um evento $E_{I'} = \{||I' \cap T| - |I' \cap M|| \geq \frac{\epsilon}{2} m\}$. A probabilidade do evento $E_{I'}$ será limitada considerando a contribuição de cada par z_{2i-1}, z_{2i} para o valor de $||I' \cap T| - |I' \cap M||$. Se z_{2i-1}, z_{2i} pertencem a I', então a contribuição é 0. Caso contrário, ou seja, se um dos elementos do par está em I' e o outro não, a contribuição é 1 com probabilidade $\frac{1}{2}$ e -1 com probabilidade $\frac{1}{2}$. Não existem mais que m pares, e daí, usando um dos Limitantes de Chernoff conclui-se:

$$\Pr(E_{I'}) \leq e^{-\epsilon^2 m/8}$$

Finalmente, pelo Teorema 2, a projeção de $\mathcal{I}$ em Z não tem mais que $(2m)^d$ intervalos. Portanto, usando o Limitante da União (Mitzenmacher e Upfal (2017)):

$$\Pr(E_2') \leq (2m)^d e^{-\epsilon^2 m/8}$$

□

Para concluir a prova deste teorema será mostrado que para

$$m \geq \frac{32d}{\epsilon^2} \ln \frac{64d}{\epsilon^2} + \frac{16}{\epsilon^2} \ln \frac{2}{\delta}$$

ocorre:

$$\Pr(E_1) \leq 2 \Pr(E_2) \leq 2 \Pr(E_2') \leq 2(2m)^d e^{-\epsilon^2 m/8} \leq \delta$$

Mas observe que:

$$2(2m)^d e^{-\epsilon^2 m/8} \leq \delta$$

Aplicando ln nos dois lados e manipulando,

$$\epsilon^2 m/8 \geq \ln(\frac{2}{\delta}) + d\ln(2m)$$

Como $m > \frac{16}{\epsilon^2}\ln\frac{2}{\delta}$ então $\epsilon^2 m/16 \geq \ln(\frac{2}{\delta})$. Agora basta provar que $\epsilon^2 m/16 \geq d\ln(2m)$. Para isso, pode-se usar o seguinte lema, demonstrado em Mitzenmacher e Upfal (2017).

Lema 5. *Se $y \geq x\ln x \geq e$, então $\frac{2y}{\ln y} \geq x$.*

E daí, aplicando o lema acima com $y = 2m \geq \frac{64d}{\epsilon^2} ln \frac{64d}{\epsilon^2}$ e $x = \frac{64d}{\epsilon^2}$:

$$\frac{4m}{\ln(2m)} \geq \frac{64d}{\epsilon^2}$$

então

$$\frac{\epsilon^2 m}{16} \geq d\ln(2m)$$

□

Portanto no exemplo da eleição, considerando que a $VCDim(E) = d$, então o instituto precisaria amostrar pelo menos $\frac{32d}{(0.01)^2}\ln(\frac{64d}{0.01}) + \frac{16}{(0.01)^2}\ln(\frac{2}{0.1})$ pessoas para ter uma pesquisa com erro de no máximo 1% com confiança de 90%.

Neste capítulo, foi visto o conceito de ϵ-amostra e, utilizando o conceito de Dimensão VC apresentado no Capítulo 3, foi demonstrado um limite no número de amostras para obtermos uma ϵ-amostra com alta confiabilidade. O Capítulo 5 utiliza esses conceitos para propor um algoritmo aleatorizado para o problema da centralidade de percolação.

5 Problema e Algoritmo Proposto

No capítulo 2 foi apresentado a centralidade de percolação, uma medida de centralidade que quantifica o impacto relativo de um vértice em uma rede durante a infestação de um vírus ou transmissão de uma doença. Além disso, foi apresentado nesse mesmo capítulo, que o algoritmo conhecido para o cálculo dessa medida é baseado no algoritmo para a centralidade de intermediação e executa em tempo $\Theta(n^3)$, onde n é o número de vértices no grafo que representa a rede.

Em Riondato e Kornaropoulos (2016), é apresentado um algoritmo baseado em amostragem para a centralidade de intermediação. Utilizando a Teoria da Dimensão VC e o conceito de ϵ-amostra, o algoritmo desenvolvido consegue estimar com alta acurácia e eficiência a medida em redes com grande número de entidades. Com base nisso, será apresentado nesse capítulo um algoritmo baseado em amostragem para a centralidade de percolação. Primeiramente, será definido o problema a ser tratado e, em seguida, definido um espaço de intervalos sobre caminhos mínimos de um grafo. Finalmente, será apresentado o algoritmo e será demonstrado sua corretude e tempo de execução.

5.1 Definição do problema a ser tratado

Como apresentado no Capítulo 2, cada vértice $v \in V(G)$ possui um valor para a centralidade de percolação. Sendo assim, o problema a ser tratado é o seguinte:

Dados: Grafo $G = (V(G), E(G))$, um estado de percolação x_v^t para todo $v \in V(G)$ no instante de tempo t, um vértice $v \in V(G)$ e $\epsilon, \delta \in (0, 1)$

Resultado: $pc^t(v) = \sum_{\substack{(u,w)\in V(G)^2 \\ u\neq v\neq w}} \frac{\sigma_{u,w}(v)}{\sigma_{u,w}} \frac{R(x_u^t - x_w^t)}{\sum_{\substack{(f,d)\in V(G)^2 \\ f\neq v\neq d}} R(x_f^t - x_d^t)}$

Para ser resolvido o problema, o mesmo será reformulado como um problema de estimativa de esperança. Em seguida, o número de amostras para termos uma estimativa com baixo erro e com alta confiabilidade, será limitado inferiormente. Isso será feito calculando a Dimensão VC do espaço de intervalos dado por X e $\mathcal{I}$ e, após isso, aplicando o Teorema 3. Por fim, será apresentado uma maneira eficiente de amostrar da distribuição π.

Na próxima seção é definido o espaço de intervalos dado por X e $\mathcal{I}$. Em seguida, na Seção 5.3, é apresentado a distribuição π e na Seção 5.4, algoritmo proposto. Por fim, a Seção 5.6 mostra uma maneira eficiente de amostrar elementos da distribuição π.

5.2 Espaço de Intervalos sobre caminhos mínimos

Nesta seção, será definido para cada vértice de um grafo G, um espaço de intervalos sobre o domínio de todos os caminhos mínimos do grafo G. Para isso, será denotado por C_G

todos os caminhos mínimos entre pares de vértices distintos de um grafo $G = (V(G), E(G))$. Logo,

$$C_G = \bigcup_{\substack{(u,w)\in V(G)^2 \\ u\neq w}} C_{u,w}$$

Agora, para cada $v \in V(G)$ o espaço de intervalos $R^v = (C_G, \mathcal{T}_v)$ é definido, onde a família de subconjuntos de C_G é dada por:

$$\mathcal{T}_v = \{c \mid c \in C_G \text{ e } v \in \text{Int}(c)\}$$

Observe então que $\mathcal{R}^v$ possui somente um intervalo, que contém todos os caminhos mínimos entre pares de vértices distintos que possuem o vértice v como vértice interno. E daí, como $|\mathcal{T}_v| = 1$, pelo Teorema 1:

$$VCDim(\mathcal{R}^v) \leq \log_2(|\mathcal{T}_v|) = \log_2(1) = 0$$

consequentemente, pela Definição 4, $VCDim(\mathcal{R}^v) \geq 0$, logo $VCDim(\mathcal{R}^v) = 0$.

Portanto, cada vértice $v \in V(G)$ define um espaço de intervalos $\mathcal{R}^v$ que tem dimensão VC igual a 0. Esse resultado será utilizado junto com o Teorema 3 para limitar o número de amostras do algoritmo proposto, apresentado na Seção 5.4.

5.3 Distribuição de probabilidades π^v

Dado um grafo $G = (V(G), E(G))$, nesta seção será definida para cada vértice $v \in V(G)$ uma distribuição de probabilidades π^v sobre C_G. Portanto, seja $v \in V(G)$ um vértice do grafo, a probabilidade de um caminho mínimo entre u e w é dada por:

$$\pi^v(p_{uw}) = \frac{R(x_u - x_w)}{\sum\limits_{\substack{(f,d)\in V(G)^2 \\ f\neq v\neq d}} R(x_f - x_d)} \left(\frac{1}{\sigma_{uw}}\right)$$

Agora, é necessário mostrar que π^v realmente é uma distribuição de probabilidade. Ou seja, mostrar que a soma das probabilidades para cada caminho mínimo em C_G é igual a 1.

Lema 6.

$$\sum_{p_{uw}\in C_G} \pi^v(p_{uw}) = 1$$

Demonstração.

$$\begin{aligned}
\sum_{p_{uw}\in C_G} \pi^v(p_{uw}) &= \sum_{p_{uw}\in C_G} \frac{R(x_u - x_w)}{\sum\limits_{\substack{(f,d)\in V(G)^2 \\ f\neq v\neq d}} R(x_f - x_d)} \frac{1}{\sigma_{uw}} \\
&= \sum_{\substack{u\in V(G) \\ u\neq v}} \sum_{\substack{w\in V(G) \\ w\neq u\neq v}} \sum_{p\in C_{uw}} \frac{R(x_u - x_w)}{\sum\limits_{\substack{(f,d)\in V(G)^2 \\ f\neq v\neq d}} R(x_f - x_d)} \frac{1}{\sigma_{uw}} \\
&= \sum_{\substack{u\in V(G) \\ u\neq v}} \sum_{\substack{w\in V(G) \\ w\neq u\neq v}} \frac{R(x_u - x_w)}{\sum\limits_{\substack{(f,d)\in V(G)^2 \\ f\neq v\neq d}} R(x_f - x_d)} \frac{\sigma_{uw}}{\sigma_{uw}} \\
&= \sum_{\substack{u\in V(G) \\ u\neq v}} \sum_{\substack{w\in V(G) \\ w\neq u\neq v}} \frac{R(x_u - x_w)}{\sum\limits_{\substack{(f,d)\in V(G)^2 \\ f\neq v\neq d}} R(x_f - x_d)} \\
&= \frac{1}{\sum\limits_{\substack{(f,d)\in V(G)^2 \\ f\neq v\neq d}} R(x_f - x_d)} \sum_{\substack{u\in V(G) \\ u\neq v}} \sum_{\substack{w\in V(G) \\ w\neq u\neq v}} R(x_u - x_w) \\
&= \frac{1}{\sum\limits_{\substack{f\in V(G) \\ f\neq v}} \sum\limits_{\substack{d\in V(G) \\ f\neq v\neq d}} R(x_f - x_d)} \sum_{\substack{u\in V(G) \\ u\neq v}} \sum_{\substack{w\in V(G) \\ w\neq u\neq v}} R(x_u - x_w) = 1
\end{aligned}$$

□

5.4 Algoritmo Proposto

No pseudo-código 2 é apresentado o algoritmo proposto. Dado um grafo G com um estado de percolação x_i, para todo $i \in V(G)$, um vértice $v \in V(G)$ e dois parâmetros $\epsilon, \delta \in (0, 1)$, onde ϵ representa o erro e δ representa a confiabilidade, a intuição do algoritmo é a seguinte. Primeiro é calculado um número r, que será o número de caminhos mínimos amostrados, usando a seguinte equação, onde $c \geq 1$ e $c \in \mathbb{R}$:

$$r = c\left(\frac{32}{\epsilon^2}\ln\frac{64}{\epsilon^2} + \frac{16}{\epsilon^2}\ln\frac{2}{\delta}\right) \tag{5.1}$$

que é suficiente para termos uma acurácia de ϵ com confiança de $1 - \delta$, como será mostrado na próxima seção. Após isso, o algoritmo repete r vezes o seguinte:

1. amostrar um vértice w com probabilidade $\frac{\sum\limits_{u\in V} R(x_u - x_w)}{\sum\limits_{\substack{(f,d)\in V(G)^2 \\ f\neq v\neq d}} R(x_f - x_d)}$.
2. dado w selecionado no passo 1, amostrar um vértice u com probabilidade $\frac{R(x_u - x_w)}{\sum\limits_{u\in V} R(x_u - x_w)}$.
3. calcular o conjunto $C_{u,w}$ dos caminhos mínimos entre u e w.
4. amostrar um caminho $p \in C_{u,w}$ de maneira uniforme.

5. adicionar $\frac{1}{r}$ na estimativa da centralidade de percolação de v se $v \in Int(p)$.

Veja que se não existe caminho entre u e w os passos 4 e 5 não serão executados, pois foi definido que quando não existe caminho então $C_{uw} = \{p_\emptyset\}$. Considere agora que S é o conjunto de todos os caminhos mínimos amostrados pelo algoritmo, então a estimativa $\tilde{pc}^t(v)$ da centralidade de percolação $pc^t(v)$ de um vértice v é a média amostral dada por:

$$\tilde{pc}^t(v) = \frac{1}{r} \sum_{p \in S} \mathbb{1}_{Int(p)}(v) = \frac{1}{r} \sum_{p \in S} \mathbb{1}_{\mathcal{T}_v}(p)$$

onde $\mathbb{1}_x(y)$ é a função indicadora definida como:

$$\mathbb{1}_x(y) = \begin{cases} 1, & \text{se } y \in x; \\ 0, & \text{caso contrário.} \end{cases}$$

A linha 6 é uma chamada ao algoritmo de Dijkstra ou de uma Busca em Largura (BFS) modificados como em Brandes (2001) para calcular, além do conjunto C_{uw}, os conjuntos P_{uv} para todo $v \in V$. Cada um desses conjuntos P_{uv} representa os predecessores do vértice v nos caminhos mínimos de u até v, onde *predecessores de* C_{uv} é definido como:

$$P_{uv} = \{s \mid s \in V(G), (s, v) \in E(G) \text{ e } d_{uv} = d_{us} + w((s, v))\}$$

Por fim, o laço da linha 9 amostra um caminho mínimo entre u e w de maneira uniforme, como visto no Lema 5 em Riondato e Kornaropoulos (2016).

Algoritmo 2: CentralidadeDePercolacao(G,x,v,ϵ,δ)

Dados: Grafo $G = (V(G), E(G))$ com $n = |V(G)|$, um estado de percolação x_i para todo $i \in V(G)$, um vértice $v \in V(G)$ e $\epsilon, \delta \in (0, 1)$.

Resultado: Uma aproximação da centralidade de percolação para o vértice $v \in V(G)$.

1 $r \leftarrow c\left(\frac{32}{\epsilon^2} \ln \frac{64}{\epsilon^2} + \frac{16}{\epsilon^2} \ln \frac{2}{\delta}\right)$;
2 $\tilde{pc}^t(v) \leftarrow 0$;
3 **para** $i \leftarrow 1$ **até** r **faça**
4 Amostre $w \in V(G)$ com probabilidade $\frac{\sum_{u \in V} R(x_u - x_w)}{\sum_{u \neq v \neq w} R(x_u - x_w)}$;
5 Amostre $u \in V(G)$ com probabilidade $\frac{R(x_u - x_w)}{\sum_{u \in V} R(x_u - x_w)}$;
6 $S_{uw} \leftarrow$ `todosCaminhosMinimos(u,w)` ;
7 **se** $S_{uw} \neq \{p_\emptyset\}$ **então**
8 $t \leftarrow w$;
9 **enquanto** $t \neq u$ **faça**
10 Amostre $z \in P_{ut}$ com probabilidade $\frac{\sigma_{uw}}{\sigma_{ut}}$;
11 **se** $z \neq u$ *e* $z = v$ **então**
12 $\tilde{pc}^t(v) \leftarrow \tilde{pc}^t(v) + \frac{1}{r}$;
13 **fim**
14 **fim**
15 **fim**
16 **fim**
17 **retorna** $\tilde{pc}^t(v)$;

5.5 Corretude

Esta seção prova que o Algoritmo 2 retorna a aproximação da centralidade de percolação para o vértice $v \in V$, tal que possui erro de no máximo ϵ com $1 - \delta$ de confiança.

Lema 7. *Com probabilidade pelo menos* $1 - \delta$*, a aproximação computada pelo Algoritmo 2 possui erro de no máximo* ϵ*.*

Demonstração. Pode-se ver que o Algoritmo 2 amostra um caminho mínimo p_{uw} com probabilidade $\pi^v(p_{uw})$, dada a maneira como os vértices u e w são selecionados, e que o laço da linhas 9-14 amostra um caminho mínimo p de maneira uniforme sobre o conjunto C_{uw}. Além disso, π^v é uma distribuição de probabilidade, como visto no Lema 6. Dado um conjunto A, será denotada por $\pi^v(A)$ a soma $\sum_{p \in A} \pi^v(p)$.

Considere agora o espaço de intervalos $\mathcal{R}^v$, como definido na Seção 5.2, e a distribuição π^v. Seja S o conjunto de caminhos mínimos amostrados pelo algoritmo. Para r como na Equação 5.1, o Teorema 3 nos diz que a amostra S é uma ϵ-amostra para $(\mathcal{R}^v, \pi^v)$ com probabilidade pelo menos $1 - \delta$. Suponha que seja esse o caso, então pela Definição 7 e pela definição de $\mathcal{R}^v$:

$$\left| \pi^v(\mathcal{T}_v) - \frac{1}{r} \sum_{p \in S} \mathbb{1}_{\mathcal{T}_v}(p) \right| = |\pi^v(\mathcal{T}_v) - \tilde{pc}^t(v)| \leq \epsilon$$

Finalmente, pela definição de π^v:

$$\pi^v(\mathcal{T}_v) = \sum_{p_{uw} \in \mathcal{T}_v} \frac{R(x_u - x_w)}{\sum\limits_{\substack{(f,d) \in V(G)^2 \\ f \neq v \neq d}} R(x_f - x_d)} \frac{1}{\sigma_{uw}} = pc^t(v)$$

Portanto,

$$\Pr(|pc^t(v) - \tilde{pc}^t(v)| \leq \epsilon) \geq 1 - \delta$$

□

5.6 Tempo de Execução

Nesta seção, será mostrado que dado um grafo $G = (V(G), E(G))$ tal $n = |V(G)|$ e $m = |E(G)|$, o algoritmo apresentado na Seção 5.4 executa em tempo $O(max(n^2, (n + m)\frac{1}{\epsilon^2} \ln \frac{1}{\delta}))$ para grafos sem peso e em tempo $O(max(n^2, (m + n \log n)\frac{1}{\epsilon^2} \ln \frac{1}{\delta})$ para grafos com peso. A análise é dividida da seguinte maneira, mostrar que, fazendo um pré-cálculo da diferença entre os estados de percolação do vértices do grafo, é possível amostrar os vértices u e w em tempo $O(n)$. Em seguida, analisar o laço da linha 9 do Algoritmo 2 e mostrar que o mesmo executa em tempo $O(m)$. Por fim, apresentar o algoritmo com as modificações necessárias e fazer a análise.

5.6.1 Amostrar u e w

O Algoritmo 3 computa a diferença dos estados de percolação entre todos os pares de vértices de um grafo. Observe, que no fim da execução, a matriz M conterá nos índices i e j a diferença entre o estado de percolação do vértice i e do vértice j, ou seja, $M[i][j] = R(x_i - x_j)$. Além disso, nos índices $n + 1$ e j esta matriz terá a soma das diferenças entre todos os vértices do

grafo e o vértice j, logo $M[n+1][j] = \sum_{i=1}^{n} R(x_i - x_j)$. E ainda, o índice $n+1$ e $n+1$, contém a soma de todas as diferenças:

$$M[n+1][n+1] = \sum_{i=1}^{n}\sum_{j=1}^{n} R(x_i - x_j) \tag{5.2}$$

Algoritmo 3: preCalculaDif(x)

Dados: O estado de percolação x_i^t para todo $i \in V(G)$
Resultado: Uma matriz M de tamanho $(n+1) \times (n+1)$.

```
1  para i ← 1 até n + 1 faça
2      para j ← 1 até n + 1 faça
3          M[i][j] ← 0 ;
4      fim
5  fim
6  para i ← 1 até n faça
7      para j ← 1 até n faça
8          dif ← (x[i] − x[j]);
9          se dif > 0 então
10             M[i][j] ← dif ;
11             M[n + 1][j] ← M[n + 1][j] + dif ;
12             M[n + 1][n + 1] ← M[n + 1][n + 1] + dif ;
13         fim
14     fim
15 fim
16 retorna M
```

Essa matriz M será usada para construir os vetores de probabilidades no processo de amostragem dos vértices u e w e o custo para construção desta matriz é $\Theta(n^2)$.

Lema 8. *O Algoritmo 3 executa em tempo* $\Theta(n^2)$.

Agora, utilizando a matriz M, é possível amostrar os vértices u e w em tempo $O(n)$. Veja que, como a posição $n+1$ e $n+1$ da matriz M contém a soma de todas as diferenças como na Equação 5.2, então basta diminuir $\sum_{u \in V(G)} (R(x_u - x_v) - R(x_v - x_u))$, para assim obter

$$\sum_{\substack{(f,d)\in V(G)^2 \\ f\neq v\neq d}} R(x_u - x_w)$$

Esse valor será utilizado para construir um vetor, sendo que a posição i do vetor contém

$$\text{prob}[i] = \frac{M[n+1][i]}{\sum\limits_{\substack{(f,d)\in V(G)^2 \\ f\neq v\neq d}} R(x_f - x_d)} = \frac{\sum\limits_{u\in V} R(x_u - x_w)}{\sum\limits_{\substack{(f,d)\in V(G)^2 \\ f\neq v\neq d}} R(x_u - x_w)}$$

como visto no Algoritmo 2. Abaixo, o algoritmo para amostrar o vértice w é apresentado. Na linha 16, o "amostraInteiroComProb" é uma chamada ao algoritmo apresentado em Vose (1991),

que amostra um inteiro com probabilidade definida por um vetor de probabilidade em tempo $O(n)$, onde n é o tamanho do vetor.

Algoritmo 4: amostraW(M,v)

Dados: Uma matriz M de tamanho $(n+1) \times (n+1)$ e um índice v.
Resultado: Um inteiro $w \in [1 \dots n]$ amostrado com probabilidade $\frac{\sum_{u \in V} R(x_u - x_w)}{\sum_{u \neq v \neq w} R(x_u - x_w)}$.

```
1  soma ← M[n + 1][n + 1]
2  para i ← 1 até n faça
3  |  se i ≠ v então
4  |  |  soma ← soma − M[i][v];
5  |  |  soma ← soma − M[v][i];
6  |  fim
7  fim
8  para i ← 1 até n faça
9  |  se i ≠ v então
10 |  |  prob[i] ← M[n+1][i] / soma;
11 |  fim
12 |  senão
13 |  |  prob[i] ← 0;
14 |  fim
15 fim
16 retorna amostraInteiroComProb (1,n,prob)
```

Utilizando também a matriz M, é possível construir o vetor de probabilidades para amostrar o vértice u, como apresentado no Algoritmo 5. O tempo de execução deste algoritmo também é $O(n)$.

Algoritmo 5: amostraU(M,v,w)

Dados: Uma matrix M de tamanho $(n+1) \times (n+1)$ e dois índices v e w.
Resultado: Um inteiro $u \in [1 \dots n]$ amostrado com probabilidade $\frac{R(x_u - x_w)}{\sum_{u \in V} R(x_u - x_w)}$.

```
1  para i ← 1 até n faça
2  |  se i ≠ v então
3  |  |  prob[i] ← M[i][w] / M[n+1][w];
4  |  fim
5  |  senão
6  |  |  prob[i] ← 0;
7  |  fim
8  fim
9  retorna amostraInteiroComProb (1,n,prob)
```

5.6.2 Amostrar um Caminho Mínimo

Agora será mostrado que é possível amostrar um caminho mínimo do conjunto C_{uw} em tempo $O(n+m)$ no caso do grafo não ter pesos e $O(m + n \log n)$ caso seja um grafo com pesos.

O trecho do Algoritmo 2 responsável por amostrar um caminho mínimo é o trecho que começa na linha 6 e termina na linha 14.

Esse trecho começa com uma chamada para o algoritmo de Dijkstra ou uma BFS, no caso do grafo ser com pesos ou sem pesos, respectivamente. O tempo de execução do algoritmo de Dijkstra utilizando uma heap de Fibonacci é de $O(m + n \log n)$ e o custo da BFS é de $O(n + m)$. Além disso, as modificações feitas nesses algoritmos para retornarem o conjunto de predecessores não altera a complexidade de tempo, como mostra Brandes (2001). Sendo assim, no lema abaixo é demonstrado que o trecho das linhas 9-14 executa em tempo $O(m)$ e, portanto, o tempo de execução para amostrar um caminho mínimo será $O(max(n + m, m)) \subseteq O(n + m)$ para grafos sem peso e $O(max(m + n \log n, m)) \subseteq O(m + n \log n)$ para grafos com peso.

Lema 9. *O trecho das linhas 9-14 do Algoritmo 2 executa em tempo $O(m)$.*

Demonstração. Primeiramente, observe que o algoritmo percorre um dos caminhos mínimos $c \in C_{uw}$. Agora, observe também que a linha 10 é uma chamada ao algoritmo de Vose (1991), que tem complexidade de tempo $O(n)$. Além disso, pela definição de predecessor, P_{uw} é um subconjunto dos vizinhos de w nos caminhos mínimos de u até w, então

$$|P_{uw}| \leq \delta_G(w)$$

Logo, seja T_{L10} o custo de execução da linha 10. Como $|P_{uw}| \leq \delta_G(w)$, então $T_{L10} \in O(\delta_G(w))$. Finalmente, como o algoritmo executa T_{L10} para cada vértice do caminho, então o custo do trecho é $O(m)$.

□

5.6.3 Algoritmo Modificado

Finalmente, será demonstrado que o Algoritmo 2 com as modificações feitas executa em tempo $O(max(n^2, (n+m)\frac{1}{\epsilon^2}\ln\frac{1}{\delta}))$ para grafos sem peso e $O(max(n^2, (m+n\log n)\frac{1}{\epsilon^2}\ln\frac{1}{\delta})$ para grafos com peso. O Algoritmo 6 apresenta o algoritmo com as modificações feitas.

Algoritmo 6: CentralidadeDePercolacao(G,x,v,ϵ,δ)

Dados: Grafo $G = (V(G), E(G))$, um estado de percolação x_i para todo $i \in V(G)$, um vértice $v \in V(G)$ e $\epsilon, \delta \in (0,1)$.

Resultado: Uma aproximação da centralidade de percolação para o vértice $v \in V(G)$.

```
1  M ← preCalculaDiferencas (x);
2  r ← c (32/ε² ln 64/ε² + 16/ε² ln 2/δ);
3  p̃c^t(v) ← 0;
4  para i ← 1 até r faça
5  |  w ← amostraW (M);
6  |  u ← amostraU (M);
7  |  S_uw ← todosCaminhosMinimos(u,w);
8  |  se S_uw ≠ {p_∅} então
9  |  |  t ← w;
10 |  |  enquanto t ≠ u faça
11 |  |  |  Amostre z ∈ P_u(t) com probabilidade σ_uw/σ_ut;
12 |  |  |  se z ≠ u e z = v então
13 |  |  |  |  p̃c^t(v) ← p̃c^t(v) + 1/r;
14 |  |  |  fim
15 |  |  fim
16 |  fim
17 fim
18 retorna p̃c^t(v)
```

Lema 10. *O Algoritmo 6 executa em tempo $O(max(n^2, (n+m)\frac{1}{\epsilon^2}\ln\frac{1}{\delta}))$ caso o grafo não tenha pesos e $O(max(n^2, (m+n\log n)\frac{1}{\epsilon^2}\ln\frac{1}{\delta})$ caso o grafo tenha pesos.*

Demonstração. Nesta prova, será chamado de T_M o tempo de execução do Algoritmo 3. Além disso, será denotado por T_{su} e T_{sw} o tempo de execução do algoritmo para amostrar u e w, respectivamente. Por fim, T_p será o tempo de execução do trecho das linhas 10- 15 e $T_{C_{uw}}$ o tempo de execução da chamada *todosCaminhosMinimos*. Portanto, como visto anteriormente:

$$\begin{aligned} T_M &= O(n^2) \\ T_{su} &= O(n) \\ T_{sw} &= O(n) \\ T_p &= O(m) \end{aligned}$$

E daí, caso o grafo não tenha peso, $T_{C_{uw}} \in O(n+m)$, portanto o custo do trecho das linhas 5- 16 é:

$$O(max(T_{su}, T_{sw}, T_{C_{uw}}, T_p))$$

ou seja,

$$O(max(n, n, (n + m), m)) \in O(n + m)$$

Logo, como esse trecho é repetido r vezes e o custo da inicialização da matriz M é $O(n^2)$, o custo total do algoritmo é:

$$O(max(n^2, (n + m)r))$$

Já para o caso do grafo com pesos, o custo de $T_{C_{uw}}$ é $O(m + n \log n)$ daí o custo das operações executadas no laço 4 é:

$$O(max(n, n, m + n \log n, m)) \in O(m + n \log n)$$

Portanto, o custo total para grafo com peso é:

$$O(max(n^2, r(m + n \log n))$$

□

Observe que o algoritmo apresentado, apesar de executar um número polinomial de operações em relação à representação do grafo G, executa um número exponencial de operações em relação a representação de ϵ. Além disso, para calcular a centralidade para todo o grafo, o algoritmo proposto e o apresentado no Capítulo 2 podem ser facilmente modificados. Porém, como o algoritmo proposto pode ter um erro maior ϵ com probabilidade δ, então utilizando o limitante da união, pode-se concluir que a probabilidade de se ter algum vértice com erro maior que ϵ no fim de n execuções escala linearmente com o número de vértices do grafo.

Apesar desses resultados, este capítulo apresentou um algoritmo baseado em amostragem para o cálculo da centralidade de percolação. O algoritmo proposto permite que o usuário escolha entre ter baixo tempo de execução ou alta precisão na estimativa, modificando os parâmetros ϵ e δ da entrada do algoritmo.

6 Conclusão

A centralidade de percolação apresentada no Capítulo 2 pode ser calculada utilizando o algoritmo exato que possui complexidade de tempo $O(n^3)$. Este trabalho apresentou um algoritmo aleatorizado baseado em amostragem para essa medida com tempo de execução $O(max(n^2, (n+m)\frac{1}{\epsilon^2}\ln\frac{1}{\delta}))$ para grafos sem peso e $O(max(n^2, (m+n\log n)\frac{1}{\epsilon^2}\ln\frac{1}{\delta})$ para grafos com peso.

Para desenvolver o algoritmo, o problema de calcular a centralidade de percolação para um vértice foi reformulado como um problema de estimar probabilidades. Para isso, foi apresentado uma distribuição de probabilidades e um espaço de intervalos de tal maneira que a centralidade de percolação de um vértice é a esperança de algum intervalo no espaço. Além disso, foi mostrado que o espaço de intervalos definido para resolver o problema possui uma Dimensão VC baixa, implicando em um baixo número de amostras necessárias para obter o resultado desejado.

O algoritmo apresentado no Capítulo 5, quando utilizado no cálculo da centralidade de todo o grafo, possui uma probabilidade de errar mais que ϵ em algum vértice que escala linearmente com o número de vértices do grafo. Como trabalhos futuros, outras técnicas e resultados da área de Complexidade de Amostra, como médias de Rademacher ou Pseudo Dimensão de Pollard, poderiam ser utilizadas com o objetivo de diminuir o número de amostras necessárias. Além disso, poderiam servir como base no desenvolvimento de um algoritmo para calcular a centralidade em todos os vértices do grafo, em que a probabilidade de existir uma entidade que tem erro maior que ϵ não escale com o número de vértices. Finalmente, uma análise experimental comparativa dos algoritmos seria útil para verificar o comportamento do algoritmo proposto na prática e como usá-lo para equilibrar tempo de execução e precisão de acordo com o desejado.

Referências

Blumer, A., Ehrenfeucht, A., Haussler, D. e Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36:929–965.

Bolfarine, H. e Bussab, W. O. (2005). *Elementos de Amostragem*, páginas 27–30. Editora Blucher.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2):163–177.

Broadbent, S. R. e Hammersley, J. M. (1957). Percolation processes: I. crystals and mazes. Em *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 53, páginas 629–641. Cambridge University Press.

Cormen, T. H., Leiserson, C. E., Rivest, R. L. e Stein, C. (2009). *Introduction to algorithms*. MIT press.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, páginas 35–41.

Haussler, D. e Welzl, E. (1986). Epsilon-nets and simplex range queries. Em *Proc. 2nd Annu. Symp. Computational geometry*, SCG '86, páginas 61–71, New York, NY, USA. ACM.

Kearns, M. J., Vazirani, U. V. e Vazirani, U. (1994). *An introduction to computational learning theory*. MIT press.

Matoušek, J. (2002). *Lectures on discrete geometry*, volume 212. Springer New York.

Mitzenmacher, M. e Upfal, E. (2017). *Probability and Computing: Randomization and Probabilistic Techniques in Algorithm and Data Analysis*. Cambridge university press.

Morettin, P. A. e Bussab, W. O. (2013). *Estatística básica*, página 151. Editora Saraiva.

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

Newman, M. E. (2002). Spread of epidemic disease on networks. *Physical review E*, 66(1):016128.

Piraveenan, M., Prokopenko, M. e Hossain, L. (2013). Percolation centrality: Quantifying graph-theoretic impact of nodes during percolation in networks. *PloS one*, 8(1):e53095.

Priberam (2013). "percolação", em dicionário priberam da língua portuguesa. `https://dicionario.priberam.org/percola%C3%A7%C3%A3o`. Acessado em 10/11/2018.

Riondato, M. e Kornaropoulos, E. M. (2016). Fast approximation of betweenness centrality through sampling. *Data Mining and Knowledge Discovery*, 30(2):438–475.

Sander, L., Warren, C., Sokolov, I., Simon, C. e Koopman, J. (2002). Percolation on heterogeneous networks as a model for epidemics. *Mathematical biosciences*, 180(1-2):293–305.

Sauer, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147.

Shelah, S. (1972). A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261.

Silva, N. N. (2015). *Amostragem Probabilística*, página 25. Editora USP.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Vapnik, V. N. e Chervonenkis, A. J. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280.

Vose, M. D. (1991). A linear algorithm for generating random numbers with a given distribution. *IEEE Transactions on software engineering*, 17(9):972–975.