Estimando Centralidade de Percolação utilizando Amostragem e Teoria da Dimensão Vapnik-Chervonenkis

Alane de Lima[1], Giovanne dos Santos[2], André Vignatti[1], Murilo da Silva[1]

[1]Departamento de Informática - Universidade Federal do Paraná (UFPR)
[2]Instituto de Matemática e Estatística - Universidade de São Paulo (USP)

Sumário

Introdução

Introdução

Centralidade em grafos de larga escala: quantifica a importância relativa de uma entidade na rede

Figure 1: Social Network Analysis Visualization (Grandjean, 2016) [3].

Características locais

Centralidade de grau

Características globais

- Centralidade de intermediação
- Centralidade de percolação

Centralidade de percolação

- Medida que generaliza a *centralidade de intermediação*

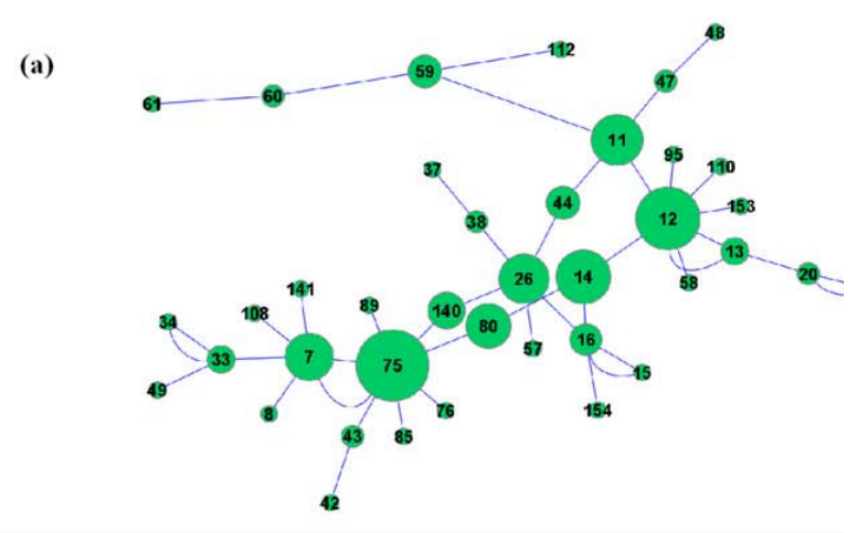


Figure 2: Centralidade de intermediação. Quanto maior o tamanho do vértice, maior sua centralidade. (Piraveenan et. al, 2013) [6]

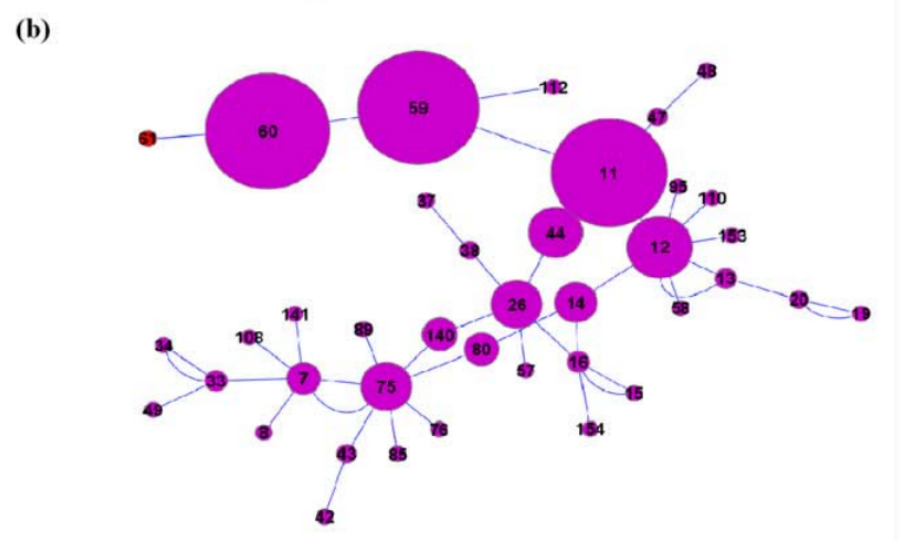


Figure 3: Centralidade de percolação para o mesmo grafo da Figura 2. Quanto maior o tamanho do vértice, maior sua centralidade. (Piraveenan et. al, 2013) [6]

Centralidade de percolação (Piraveenan et. al (2013))

- Processo de infestação da rede (doenças, *fake news*)

Centralidade de percolação (Piraveenan et. al (2013))

- Processo de infestação da rede (doenças, *fake news*)
- (Broadbent e Hammersley, 1957) [2] ⇒ passagem de fluido na rede; cada nó da rede possui um **estado de percolação** e cada parte do "meio" possui uma probabilidade de transmitir o fluido ou não.

Algoritmos exatos

- Centralidade de intermediação: $\mathcal{O}(nm)$ (Brandes, 2001) [1])
- Centralidade de percolação: $\mathcal{O}(n^3)$ (Floyd-Warshall)

Algoritmos exatos

- Centralidade de intermediação: $\mathcal{O}(nm)$ (Brandes, 2001) [1])
- Centralidade de percolação: $\mathcal{O}(n^3)$ (Floyd-Warshall)

Por que usar um algoritmo de aproximação?

- Na prática, os algoritmos citados são muito custosos em grafos de larga escala.

Algoritmos exatos

- Centralidade de intermediação: $\mathcal{O}(nm)$ (Brandes, 2001) [1])
- Centralidade de percolação: $\mathcal{O}(n^3)$ (Floyd-Warshall)

Por que usar um algoritmo de aproximação?

- Na prática, os algoritmos citados são muito custosos em grafos de larga escala.
- Estimativas que respeitam parâmetros de confiança e qualidade desejados podem ser suficientes.

Trabalho em andamento

Proposta: Algoritmo aleatorizado de aproximação

Estimar a centralidade de percolação dos vértices de um grafo $G = (V, E)$, dados parâmetros de qualidade e confiança, respectivamente, $\epsilon, \delta \in [0, 1]$.

Trabalho em andamento

Proposta: Algoritmo aleatorizado de aproximação

Estimar a centralidade de percolação dos vértices de um grafo $G = (V, E)$, dados parâmetros de qualidade e confiança, respectivamente, $\epsilon, \delta \in [0, 1]$.

Base do algoritmo de aproximação:

Trabalho em andamento

Proposta: Algoritmo aleatorizado de aproximação

Estimar a centralidade de percolação dos vértices de um grafo $G = (V, E)$, dados parâmetros de qualidade e confiança, respectivamente, $\epsilon, \delta \in [0, 1]$.

Base do algoritmo de aproximação:

- Abordagem de (Riondatto e Kornaropoulos, 2016) [7] (centralidade de intermediação)

Trabalho em andamento

Proposta: Algoritmo aleatorizado de aproximação

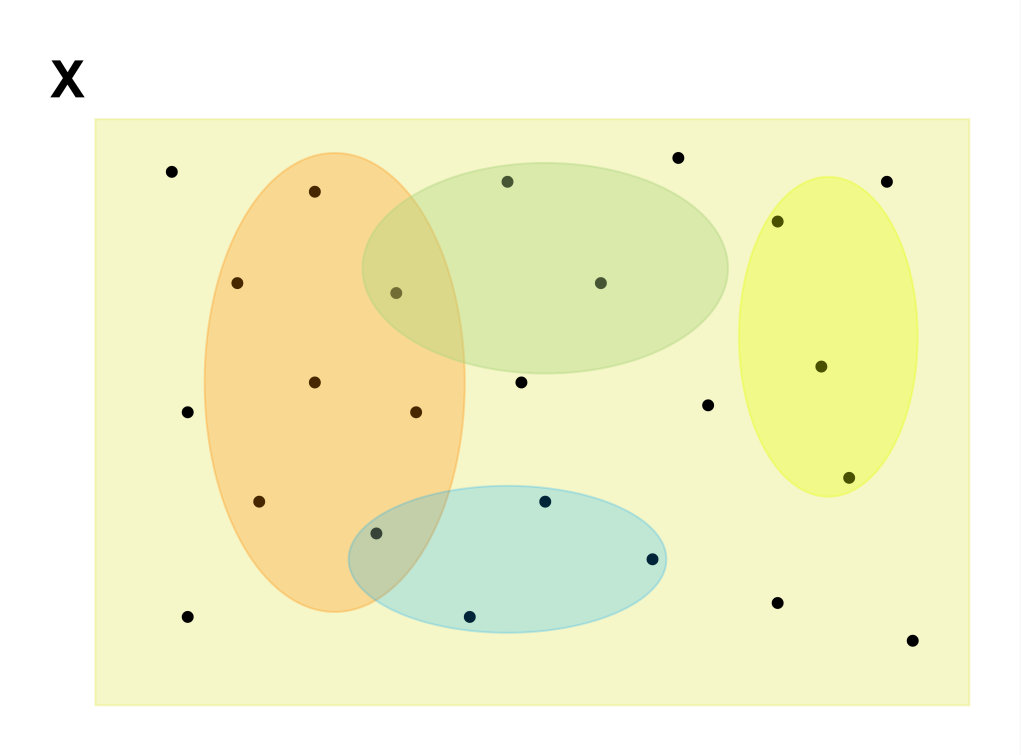
Estimar a centralidade de percolação dos vértices de um grafo $G = (V, E)$, dados parâmetros de qualidade e confiança, respectivamente, $\epsilon, \delta \in [0, 1]$.

Base do algoritmo de aproximação:

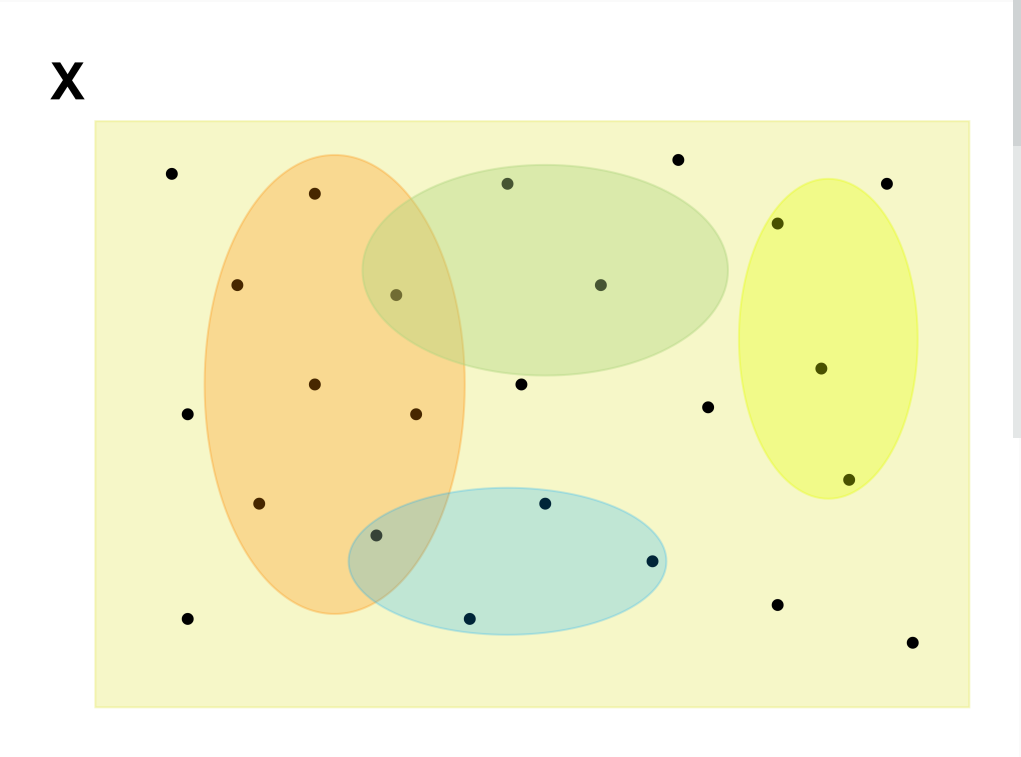
- Abordagem de (Riondatto e Kornaropoulos, 2016) [7] (centralidade de intermediação)
- Complexidade de Amostra: Teoria de Pseudodimension e ϵ-amostra

Preliminares Matemáticos

Complexidade de amostra: VC-Dimension



Complexidade de amostra: VC-Dimension

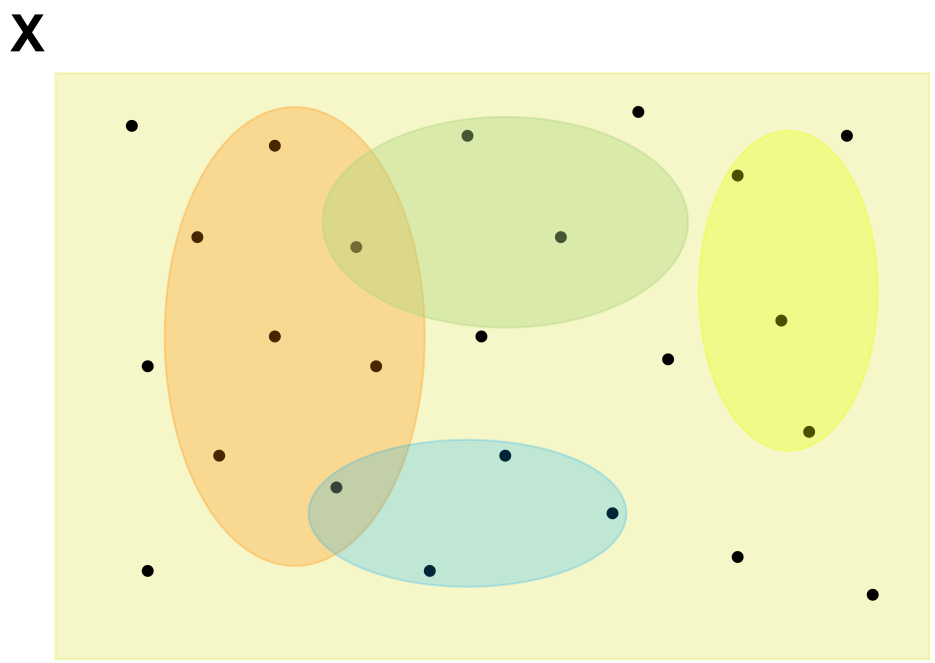


Espaço de intervalos

$R = (X, \mathcal{I})$

- Domínio X
- Conjunto de intervalos $\mathcal{I}$ (subconjuntos de X)

Complexidade de amostra: VC-Dimension



Espaço de intervalos

$R = (X, \mathcal{I})$

- Domínio X
- Conjunto de intervalos $\mathcal{I}$ (subconjuntos de X)

Projeção dos intervalos em um subconjunto S ⊆ X

$\mathcal{I}_S = \{S \cap I, \text{ tal que } I \in \mathcal{I}\}$

VC-Dimension(R)

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

ϵ-amostra

Sejam $R = (X, \mathcal{I})$ um espaço de intervalos, e $\mathcal{D}$ 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 $I \in \mathcal{I}$,

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

onde $\Pr_{\mathcal{D}}(I)$ é a probabilidade de um elemento pertencer ao intervalo I de acordo com a distribuição $\mathcal{D}$.

ϵ-amostra

Teorema ϵ-amostra (prova em [Li et. al, 2001) [4]]

Sejam $R = (X, \mathcal{I})$ um espaço de intervalos tal que $VCDim(R) \leq d$ e $\mathcal{D}$ uma distribuição de probabilidade em X. Para qualquer $\epsilon, \delta \in [0, 1]$, existe

$$r = \frac{c}{\epsilon^2}\left(d + \ln\frac{1}{\delta}\right)$$

tal que uma amostra aleatória com respeito a $\mathcal{D}$ de tamanho r é uma ϵ-amostra para X com probabilidade $1 - \delta$, para constante $c > 0$.

Estima-se que $c = 0.5$ [Löffler e Phillips, 2009) [5]].

Algoritmo para centralidade de intermediação

Centralidade de intermediação b(v)

Definição

$$b(v) = \frac{1}{n(n-1)} \sum_{\substack{(u,w)\in V^2 \\ u\neq w}} \frac{\sigma_{uw}(v)}{\sigma_{uw}}$$

dado um grafo $G = (V, E)$ e vértice $v \in V$. Na definição, σ_{uw} é a quantidade de caminhos mínimos de u a w e $\sigma_{uw}(v)$ é a quantidade de caminhos mínimos de u a w que passam por v.

Espaço de intervalos

Espaço de intervalos $R = (X, \mathcal{I})$

- Domínio $X = S_G$ (caminhos mínimos de $G = (V, E)$)
- Intervalos: $\tau_V = \{p \in S_G, \text{ tal que } v \in Int(p)\}, \forall v \in V$
- Distribuição de probabilidades $\mathcal{D}(p_{uw}) = \frac{1}{n(n-1)} \frac{1}{\sigma_{uw}}$

Teorema

Seja $diam(G)$ o diâmetro do grafo G. Então,

$$VCDim(R) \leq \lfloor \lg diam(G) - 2 \rfloor + 1.$$

O que estamos calculando?

Seja $\tilde{b}(v)$ o valor aproximado da centralidade de intermediação e seja $S \subseteq X$ amostra de tamanho r.

Temos que provar que $|b(v) - \tilde{b}(v)| = \left|\Pr_{\mathcal{D}}(\tau_v) - \frac{|S \cap \tau_v|}{|S|}\right|$, pois

$$\begin{aligned}
\Pr_{\mathcal{D}}(\tau_v) &= \sum_{p_{uw} \in \tau_v} \mathbb{1}_{\tau_v}(p_{uw}) \frac{1}{n(n-1)} \frac{1}{\sigma_{uw}} \\
&= \frac{1}{n(n-1)} \sum_{\substack{(u,w) \in V^2 \\ u \neq v \neq w}} \sum_{p \in S_{uw}} \mathbb{1}_{\tau_v}(p) \frac{1}{\sigma_{uw}} \\
&= \frac{1}{n(n-1)} \sum_{\substack{(u,w) \in V^2 \\ u \neq v \neq w}} \frac{\sigma_{uw}(v)}{\sigma_{uw}} = b(v).
\end{aligned}$$

$$\tilde{b}(v) = \frac{1}{r} \sum_{p_{uw} \in S} \mathbb{1}_{\tau_v}(p_{uw}) = \frac{|S \cap \tau_v|}{|S|}.$$

O que estamos calculando?

Utilizando VCDim(R) e o teorema da ϵ-amostra:

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

para tamanho de amostra

$$r = \frac{c}{\epsilon^2}\left(\lfloor \lg diam(G) - 2 \rfloor + 1 + \ln \frac{1}{\delta}\right)$$

O valor de $diam(G)$ é obtido por meio de uma 2-aproximação, proposta por [Riondatto e Kornaropoulos (2016)], que executa em $\mathcal{O}(m \log n)$ ou $\mathcal{O}(n + m)$, caso o grafo tenha ou não pesos.

Algoritmo para centralidade de percolação

Centralidade de percolação p(v) para todo $v \in V$

Definição

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

dado um grafo $G = (V, E)$, vértice $v \in V$, os estados de percolação x_v, $\forall v \in V$ e $R(x) = \max\{x, 0\}$. Na definição, σ_{uw} é a quantidade de caminhos mínimos de u a w e $\sigma_{uw}(v)$ é a quantidade de caminhos mínimos de u a w que passam por v.

Centralidade de percolação p(v)

Espaço de intervalos $R = (X, \mathcal{I})$

- Domínio $X = S_G$ (caminhos mínimos de $G = (V, E)$)
- Conjuntos: $\tau_v = \{p \in S_G, \text{ tal que } v \in Int(p)\}, \forall v \in V$
- Distribuição de probabilidades $\mathcal{D}(p_{uw}) = \frac{1}{n(n-1)} \frac{1}{\sigma_{uw}}$
- **Família de funções $\mathcal{F}$**:

$$f_v(p_{uw}) = \frac{R(x_u - x_w)}{\sum_{(f,d) \in V^2 : f \neq v \neq d} R(x_f - x_d)} \mathbb{1}_{\tau_v}(p_{uw}).$$

Dificuldade de utilizar VC-Dimension aqui: funções que retornam valores em $[0, 1]$. Portanto, utilizamos a teoria da *pseudo-dimension*, que generaliza VC-Dimension. Apesar de mais complicado, conseguimos fazer para todos os vértices ☺.

Centralidade de percolação p(v) para um vértice

Quando o problema é restrito ao cálculo da estimativa da centralidade de percolação para *um vértice específico*, VC-Dimension se aplica, pois um par de vértices $(u, w) \in V^2$ pode ser amostrado com probabilidade

$$\frac{R(x_u - x_w)}{\sum_{(f,d)\in V^2 : f \neq v \neq d} R(x_f - x_d)},$$

e assim,

$$\tilde{p}(v) = \frac{1}{r} \sum_{p_{uw} \in S} \mathbb{1}_{\tau_v}(p_{uw}) = \frac{|S \cap \tau_v|}{|S|}.$$

Centralidade de percolação p(v) para um vértice

Teorema.

Seja $R = (S_G, \mathcal{I})$, onde $\mathcal{I} = \{\tau_V\}$. Então $VCDim(R) = 0$.

Utilizando VCDim(R) e o teorema da ϵ-amostra:

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

para tamanho de amostra

$$r = \frac{c}{\epsilon^2} \ln \frac{1}{\delta}$$

Algoritmo proposto neste trabalho

Algoritmo 1: CENTRALIDADEDEPERCOLACAO(G,x,v,ϵ,δ)

Entrada: Grafo $G = (V, E)$ com $n = |V|$, estados de percolação x, vértice $v \in V$, erro ϵ, confiança δ.

Saída: Aproximação $\tilde{p}(v)$ para a centralidade de percolação do vértice v.

1 $\tilde{p}(v) \leftarrow 0, \forall v \in V, \quad r \leftarrow \frac{c}{\epsilon^2} \ln \frac{1}{\delta}$;

2 **para** *cada* $i \leftarrow 1$ *até* r **faça**

3 amostre w com probabilidade $\frac{\sum_{u \in V} R(x_u - x_w)}{\sum_{u \neq v \neq w} R(x_u - x_w)}$ e u com probabilidade $\frac{R(x_u - x_w)}{\sum_{u \in V} R(x_u - x_w)}$;

4 $S_{uw} \leftarrow$ TODOSCAMINHOSMÍNIMOS(u, w);

5 $p_{uw} \leftarrow$ caminho mínimo amostrado uniformemente em S_{uw};

6 $\tilde{p}(v) \leftarrow \tilde{p}(v) + 1/r$, se $v \in Int(p_{uw})$;

7 **retorna** $\tilde{p}(v)$

O algoritmo que pré-computa os valores $R(x_f - x_d)$, para todo $(f, d) \in V^2$, executa em tempo $\mathcal{O}(n \log n)$.

Tempo de execução

A complexidade de tempo do algoritmo para a estimativa da centralidade de percolação $\tilde{p}(v)$, para todo $v \in V$, é

- $\mathcal{O}((n+m)r)$ para grafos sem peso

Tempo de execução

A complexidade de tempo do algoritmo para a estimativa da centralidade de percolação $\tilde{p}(v)$, para todo $v \in V$, é

- $\mathcal{O}((n+m)r)$ para grafos sem peso
- $\mathcal{O}((m \log n)r)$ para grafos com peso

Conclusão

Considerações Finais

- Algoritmo para estimar a centralidade de percolação.

Considerações Finais

- Algoritmo para estimar a centralidade de percolação.
- Diferença na aproximação para o problema de um vértice versus problema para todos os vértices.

Considerações Finais

- Algoritmo para estimar a centralidade de percolação.
- Diferença na aproximação para o problema de um vértice versus problema para todos os vértices.
- Como trabalhos futuros $\Rightarrow$ análise experimental.

References i


U. Brandes.
A faster algorithm for betweenness centrality.
Journal of Mathematical Sociology, 25(163):163–177, 2001.


S. Broadbent and J. M. Hammersley.
Percolation processes: I. crystals and mazes.
Math. Proc. of the Cambridge Philosophical Society, 53(3):629–641, 1957.


M. Grandjean.
La connaissance est un réseau.
Les cahiers du numérique, 10(3):37–54, 2014.


Y. Li, P. M. Long, and A. Srinivasan.
Improved bounds on the sample complexity of learning.
Journal of Computer and System Sciences, 62(3):516–527, 2001.

References ii


M. Löffler and J. M. Phillips.
Shape fitting on point sets with probability distributions.
In *European Symposium on Algorithms*, pages 313–324. Springer, 2009.


M. Piraveenan, M. Prokopenko, and L. Hossain.
Percolation centrality: Quantifying graph-theoretic impact of nodes during percolation in networks.
PLOS ONE, 8(1):1–14, 2013.


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