International Conference in Data Mining (ICDM 2025)

Fast Percolation Centrality Approximation with Importance Sampling

Aalto University School of Science (★)

(†)

Antonio Cruciani★

Leonardo Pellegrina†

Our Paper

Abstract

Percolation Centrality is a useful measure to quantify the importance of the vertices in a contagious process or to diffuse information. However, it is **impractical** to compute the exact percolation centrality on modern-sized networks.

- There are **key limitations of state-of-the-art** sampling-based approximation algorithms
- We show that, in most cases, the SOTA cannot achieve accurate solutions efficiently

- We propose **PercIS** a sampling algorithm based on Importance Sampling
- PercIS severely overperforms the SOTA, both, theoretically and experimentally.

Problem Statement

Input: A graph $G = (V, E)$ with $n = |V|$ and $m = |E|$, and percolation states $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in [0, 1]^n$

Problem: Compute the *exact* **percolation centrality** for each node v,

$$p(v) = \sum_{s \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \cdot \kappa(s, t, v) \in [0, 1]$$

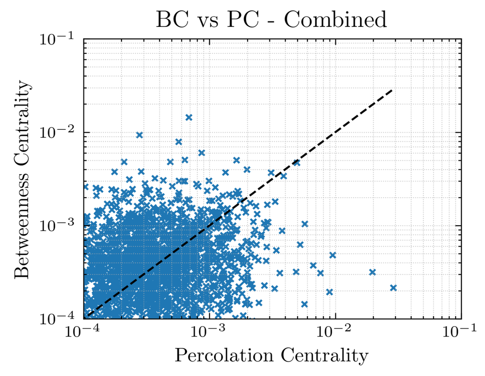
- $\sigma_{st}(v)$ number of shortest paths between s and t passing through v
- σ_{st} overall number of shortest paths between s and t
- $\kappa(s, t, v) = \frac{R(x_s - x_t)}{\sum_{u \neq v \neq w} R(x_u - x_w)}$
- $R(x) = \max(0, x)$

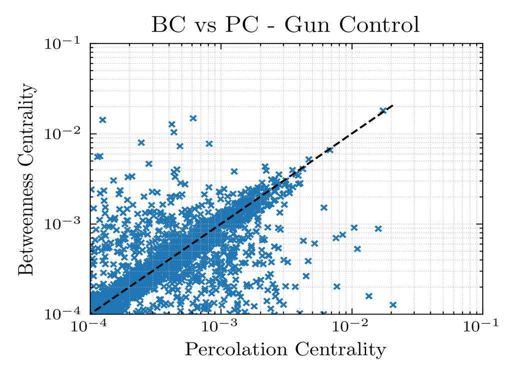
Challenge: Exact computation requires $O(n \cdot m)$ time!

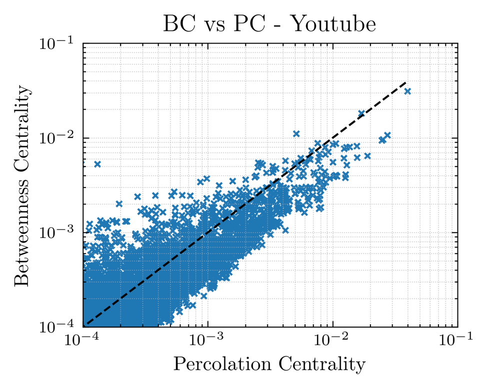
Goal: Compute an ε-approximation of the percolation centrality:

$$|p(v) - \tilde{p}(v)| \le \varepsilon, \quad \forall v \in V$$

Use case: information/contagion propagation in networks







	Jaccard Similarity Top-K		
Graph	**10**	**50**	**100**
Guns	0.053	0.087	0.117
Combined	0.0	0.031	0.015
Youtube	0.429	0.369	0.504

Jaccard similarity between betweenness and percolation centrality rankings.

State of the art

Lima et al. [1,2] generalised the techniques for the Betweenness centrality to the Percolation centrality.

High level idea:

- Randomly sample shortest paths of the graph
- Use the (weighted) fraction of the paths that traverse v as an estimate of its percolation centrality.

Cons: Technical issues that prevent these methods to be useful in practical applications.

No truly effective algorithm exists to approximate the percolation centrality.

Our Approach: Importance Sampling

Distribution: We define $\tilde{\kappa} : V \times V \to [0, 1]$

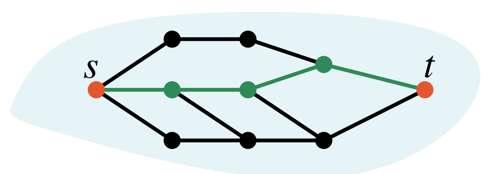
$$\tilde{\kappa}(s, t) = \frac{R(x_s - x_t)}{\sum_{u \neq w} R(x_u - x_w)}$$

For any shortest path τ_{st}, we consider the *importance distribution* :

$$q(\tau_{st}) = \frac{\tilde{\kappa}(s, t)}{\sigma_{st}}$$

Sampling from q

(1) Sample two nodes s and t with probability $\tilde{\kappa}(s, t)$;
(2) Compute the set of shortest paths Γ_{st} from s to t;
(3) Choose one shortest path uniformly at random from Γ_{st}.



The estimator and its properties

Let $\mathcal{S} = \{\tau^1, \tau^2, \ldots, \tau^\ell\}$ be a sample of ℓ i.i.d. shortest paths from q.

$$\tilde{p}(v) = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{\kappa(s, t, v)}{\tilde{\kappa}(s, t)} \mathbb{1}\left[v \in \mathcal{I}(\tau_{st}^i)\right]$$

- The estimator is *unbiased*.
- The variance is bounded by $\mathrm{Var}_q[\tilde{p}(v)] \le \hat{d}p(v)$

Where $\hat{d}$ is the *likelihood ratio*

$$\hat{d} = \max_{v \in V} \left\{ \max_{\substack{s,t \in V, \\ \tilde{\kappa}(s,t) > 0}} \frac{\kappa(s, t, v)}{\tilde{\kappa}(s, t)} \right\}$$

PercIS

Algorithm 1: PercIS

Input: Graph $G = (V, E)$, percolation states $x_1, x_2, \ldots, x_n$, $\ell_1 \ge 2$, $\varepsilon, \delta \in (0, 1)$.
Output: ε-approximation of $\{p(v), v \in V\}$ with probability $\ge 1 - \delta$

```
D ← VertexDiamUB(G);
S ← ImportanceSampler(G, {x_v}, ℓ_1);
forall v ∈ V do p̃(v) ← (1/ℓ) Σ_{i=1}^{ℓ} κ(s,t,v)/κ̃(s,t) 𝟙[v ∈ I(τ^i_st)]
ρ̂ ← ρ̃(S) + sqrt( 2Λ(S) log(8/δ) / ℓ_1 ) + 7D log(8/δ) / (3(ℓ_1 − 1));
v̂ ← d̂² max_{v∈V} { p̃(v) + sqrt( 2p̃(v) log(4/δ) / ℓ_1 ) + log(4/δ) / (3ℓ_1) };
x̂ ← d̂/2 − sqrt( d̂²/4 − min{d̂²/4, v̂} );
ℓ ← sup_{x∈(0,x̂]} { d̂² ln(4d̂ρ̂/(xδ)) / ( g(x) h(εd̂/g(x)) ) };
S ← ImportanceSampler(G, {x_v}, ℓ);
forall v ∈ V do p̃(v) ← (1/ℓ) Σ_{i=1}^{ℓ} κ(s,t,v)/κ̃(s,t) 𝟙[v ∈ I(τ^i_st)]
return {p̃(v), v ∈ V}
```

Theoretical guarantees of PercIS

ImportanceSampler draws ℓ samples from q in time $O(n + \ell(\log n + T_{BBFS}))$ and space $O(n + m)$.

Define $\hat{v}$ and $\hat{\rho}$ such that

avg. path length!

$$\max_{v \in V} \mathrm{Var}_q[\tilde{p}(v)] \le \hat{v}, \qquad \sum_{v \in V} p(v) \le \hat{d}\hat{\rho}$$

Given a sample $\mathcal{S} = \{\tau^1, \ldots, \tau^\ell\}$ of ℓ shortest paths sampled from q, and $\delta, \varepsilon \in (0, 1)$ then

$$\ell \approx \frac{\left(2\hat{v} + \frac{2}{3}\varepsilon\hat{d}\right)}{\varepsilon^2} \left(\ln(\hat{d}\hat{\rho}/\hat{v}) + \ln(2/\delta)\right)$$

gives an ε-approximation of the percolation centrality with probability $\ge 1 - \delta$

PercIS vs UNIF

State Gap: $$\Delta = \min_{v \in V} \max_{s \neq v \neq t} (x_s - x_t)$$

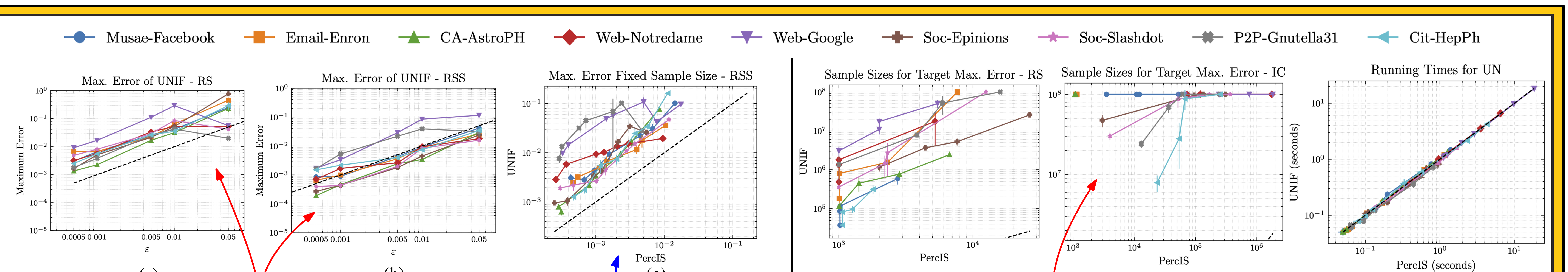
- When $\Delta \in \Omega(1)$, the likelihood ratio $\hat{d}$ of the IS distribution q is $\hat{d} \in O(1)$
- There exists instances with $\Delta \in \Omega(1)$ where the likelihood ratio of the uniform distribution is $\Omega(n)$
- There exists instances with $\Delta \in \Omega(1)$ where at least $\Omega(n^2)$ random samples are needed by UNIF, while $O(n)$ random samples are sufficient for PercIS

For all the considered real world networks, it holds $\Delta = 1$

Networks and Experiments

Graph	$\lvert V \rvert$	$\lvert E \rvert$	D	ρ	Type
P2P-Gnutella31	62586	147892	31	7.199	D
Cit-HepPh	34546	421534	49	5.901	D
Soc-Epinions	75879	508837	16	2.755	D
Soc-Slashdot	82168	870161	13	2.135	D
Web-Notredame	325729	1469679	93	9.265	D
Web-Google	875713	5105039	51	9.713	D
Musae-Facebook	22470	170823	15	2.974	U
Email-Enron	36692	183831	13	2.025	U
CA-AstroPH	18771	198050	14	2.194	U

- **Random Seeds (RS):** small number of nodes with $x_v = 1$ and the rest to 0
- **Random Seeds Spread (RSS):** Simulation of infection spreading from random seeds
- **Isolated Component (IC):** small isolated component with some nodes $x_v = 1$ and the rest to 0
- **Uniform States (UN):** $x_v \sim \mathrm{Uniform}([0, 1])$



Not a valid ε-apx!!

PercIS consistently outperforms UNIF

UNIF fails entirely!!

PercIS has a running time comparable to UNIF

(a-b) Maximum Error of UNIF on random samples of size $O(\log(D/\delta)/\varepsilon^2)$ (bound in [1]). (c) Maximum Errors of PERCIS (x axes) and UNIF (y axes) on random samples of fixed sizes $\ell \in [10^3, 10^6]$

(d-e) Sample sizes required to obtain a Maximum Error $\le \varepsilon$ PercIS (x-axes) and UNIF (y-axes). (f) Comparison between the running times (in seconds) of UNIF and PERCIS on fixed sample sizes $\ell \in [10^3, 10^6]$.

[1] A. M. de Lima, M. V. da Silva, and A. L. Vignatti, "Estimating the percolation centrality of large networks through pseudo-dimension theory," in ACM KDD, 2020.

[2] A. M. De Lima, M. V. Da Silva, and A. L. Vignatti, "Percolation centrality via rademacher complexity," Discr. Appl. Mathematics, 2022.

Paper ID: DM681