

Understanding Influence Maximization via Higher-Order Decomposition

Zonghan Zhang, Zhiqian Chen
Computer Science and Engineering
Mississippi State University



MISSISSIPPI STATE
UNIVERSITY™

Department of Computer Science
and Engineering

Agenda

- Motivation
- Introduction to **Sobol** Indices
- Influence Decomposition with **Sobol** Indices
- Proposed Method for IM
- Empirical Experiment

Motivation

- **Influence Maximization**
 - find k seeds that maximizes the influence spread

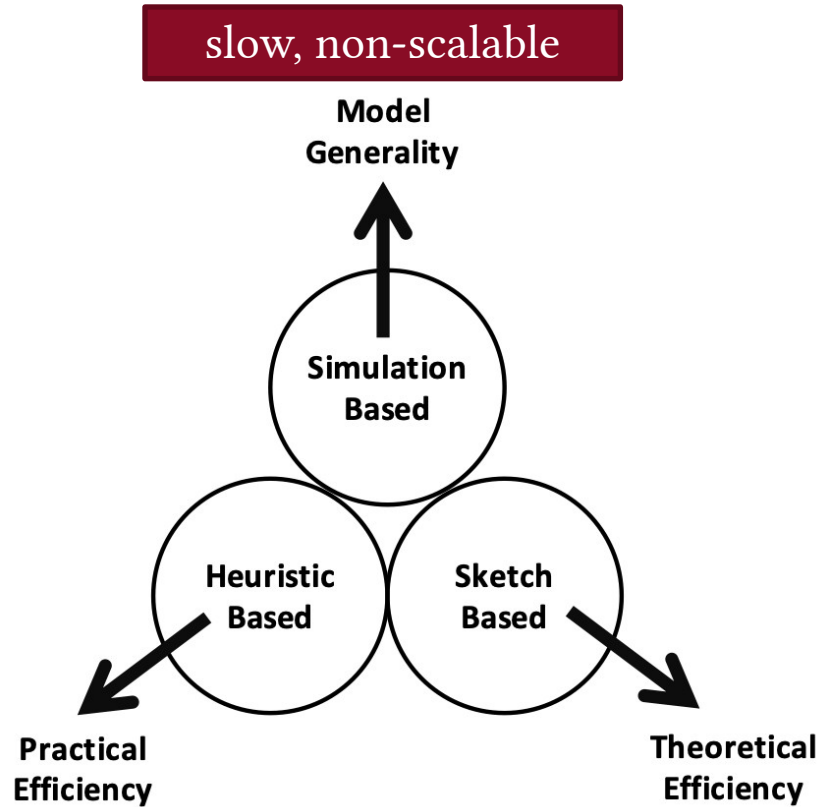


Motivation

- Application



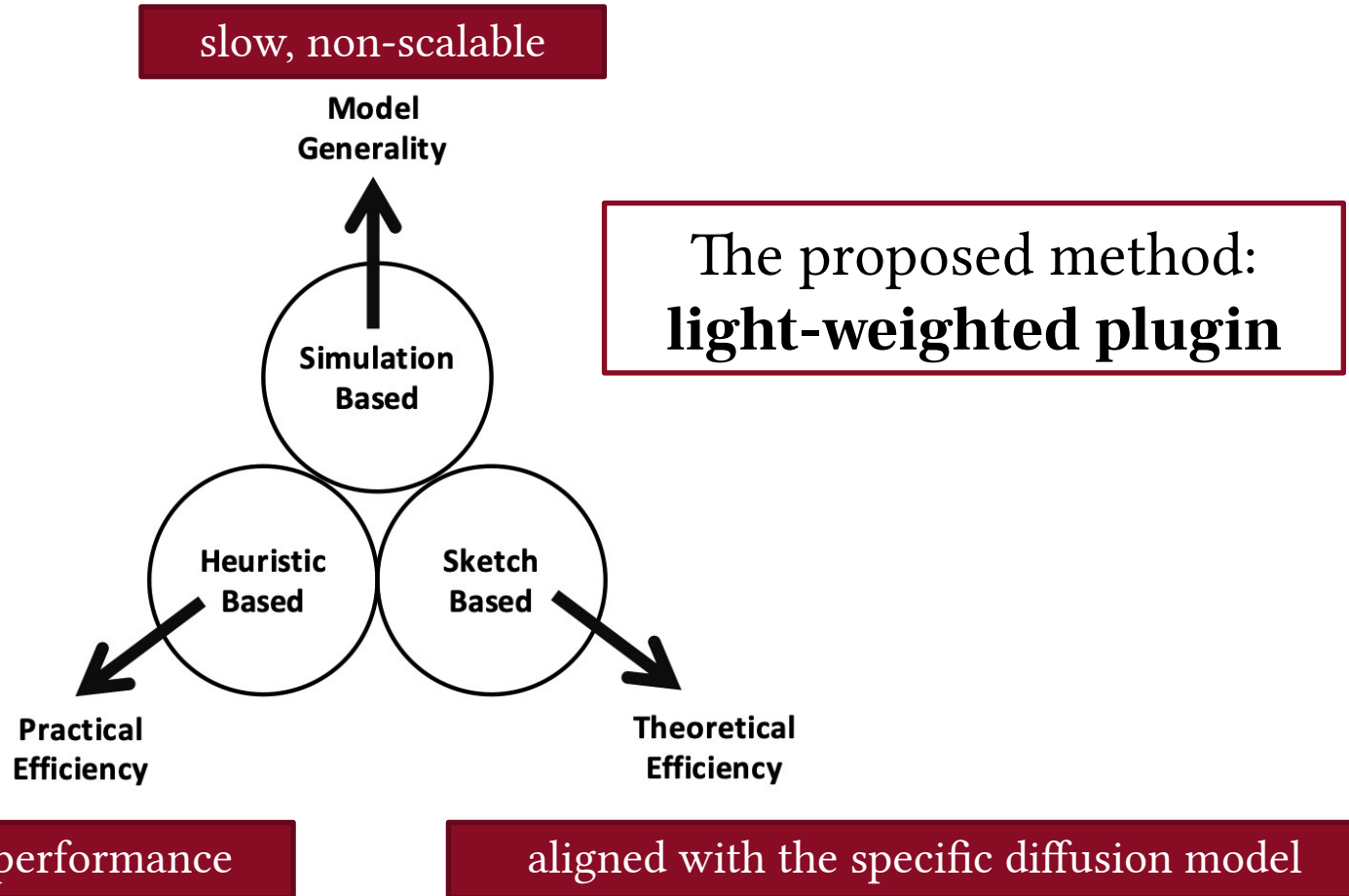
Motivation



cannot guarantee performance

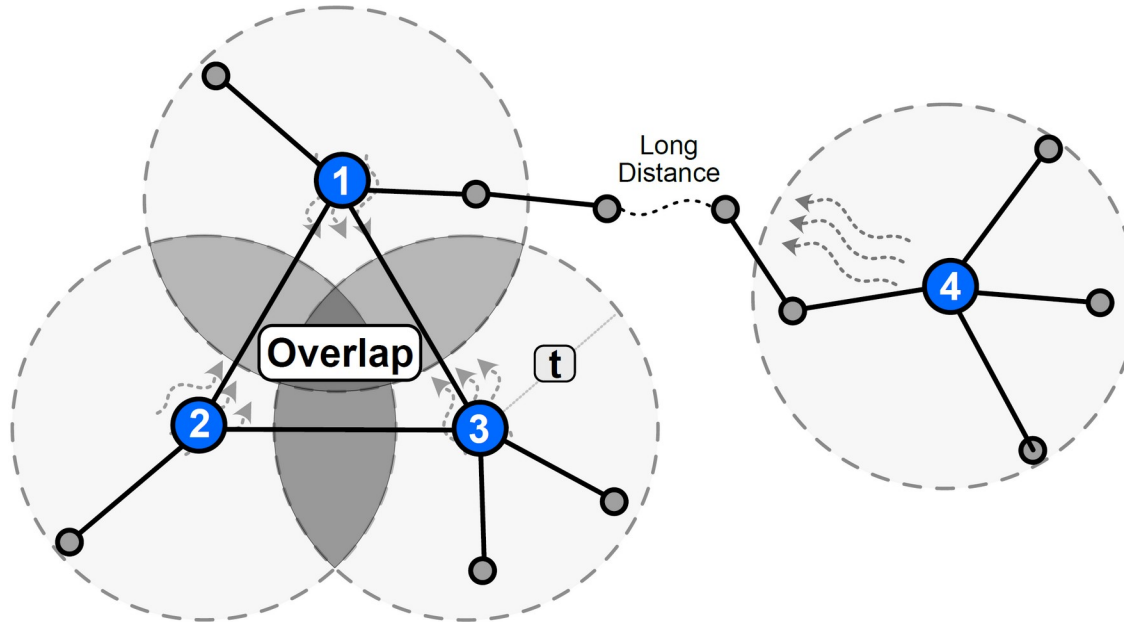
aligned with the specific diffusion model

Motivation



Motivation

Select 4 seed nodes to spread

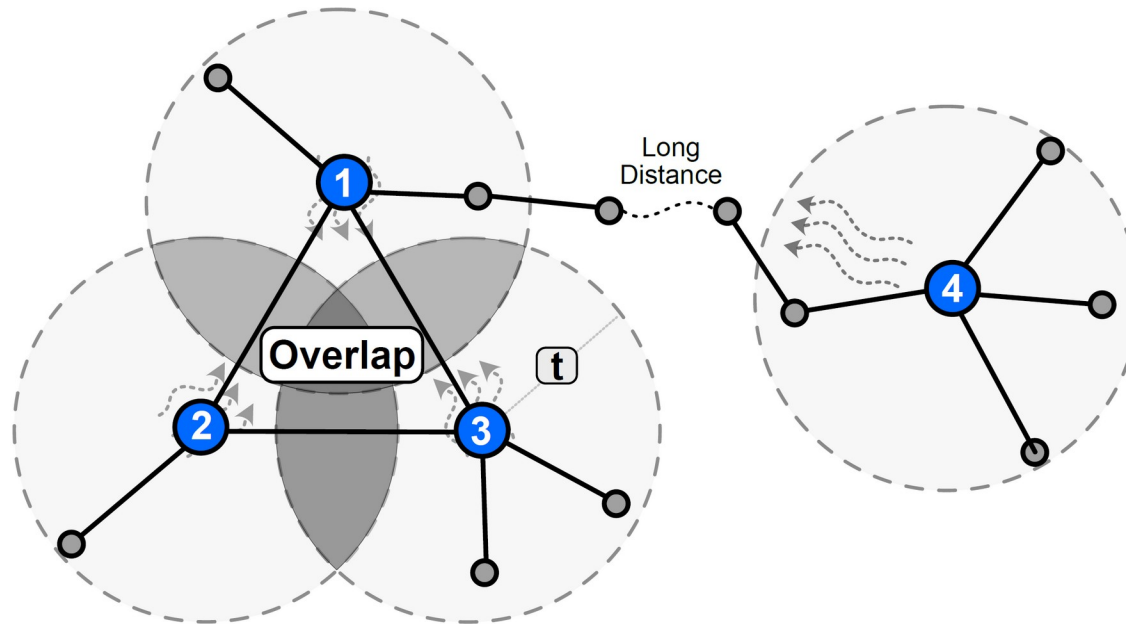


Observation 1: Nodes 1, 2, 3 have high interactions among them.

Observation 2: Node 4 has little interaction with the other nodes.

Motivation

Select 4 seed nodes to spread

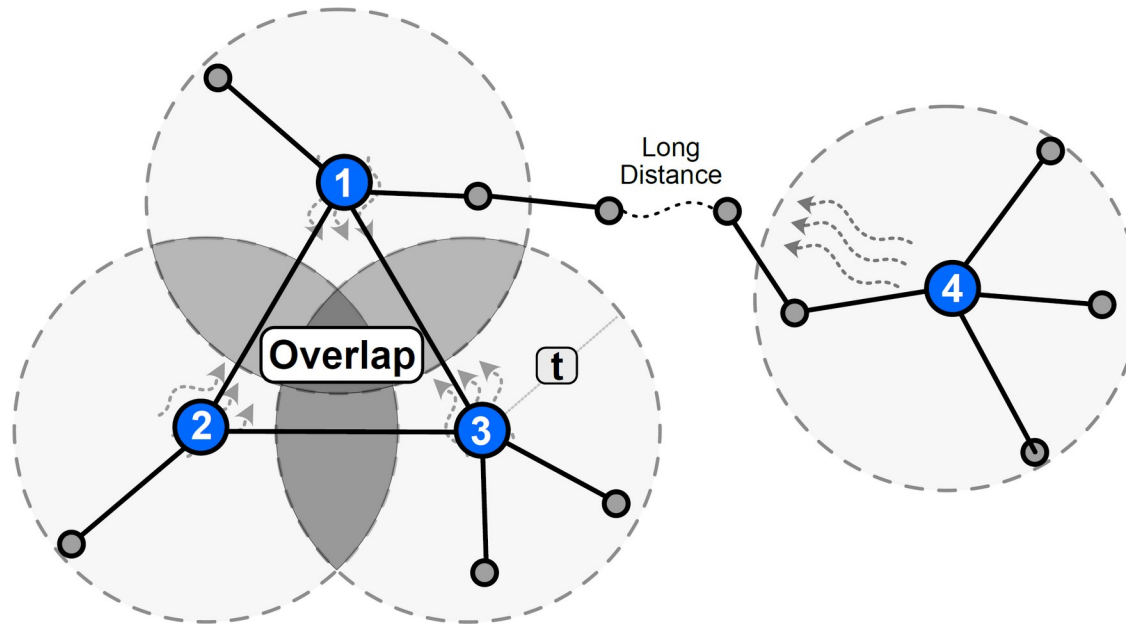


Related work:

- Local methods such as LIME and Grad-CAM ignore global effect.
- Marginal contribution fails to distinguish first- and higher-order effect.
- **Shapley value** will be discussed later.

Motivation

Select 4 seed nodes to spread



Related work

- Local method
- Marginal contribution
- Shapley value

The proposed method:
**Sobol Indices for
Higher-order Decomposition**

effect.
-order effect.

will be discussed later.

Problem Setup

- Input
 - Un-directional graph $G = (V, E)$
 - A budget k
- Output
 - A k -sized seed set Ω that maximize the influence spread $\sigma(\Omega)$
- Goal
 - Explain the contribution of a seed from a **higher-order** perspective
 - Design an **efficient and effective** IM schema

Agenda

- Motivation
- Introduction to **Sobol** Indices
- Influence Decomposition with **Sobol** Indices
- Proposed Method for IM
- Empirical Experiment

What is Sobol Index?

- Analysis of Variance-based (ANOVA) Method
- Decompose a function into different orders.

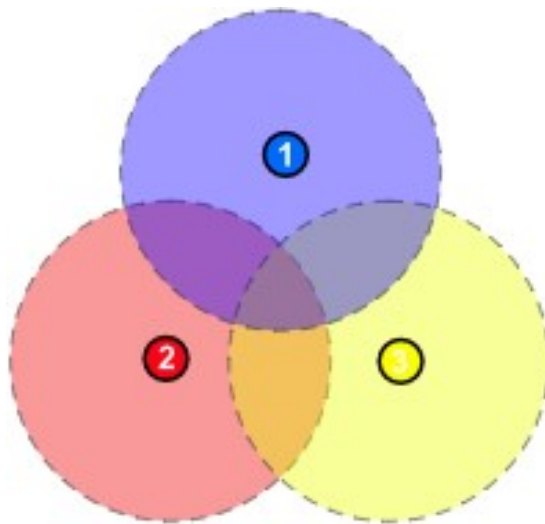
$$Y = f_0 + \overbrace{\sum_i f_i}^{\text{first-order}} + \overbrace{\sum_i \sum_{j>i} f_{ij} + \dots + f_{12\dots n}}^{\text{higher-order}}$$

What is Sobol Index?

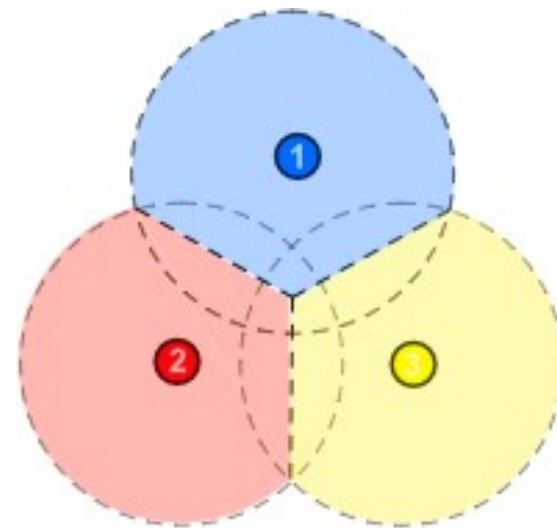
- First-order effect:
 - Variance accounted for one single variable only.
- Higher-order effect:
 - Variance brought by the interactions among variables.
- Total effect:
 - Combines first- and higher-order effects

Comparing with Shapley

- Why Sobol indices work better here?



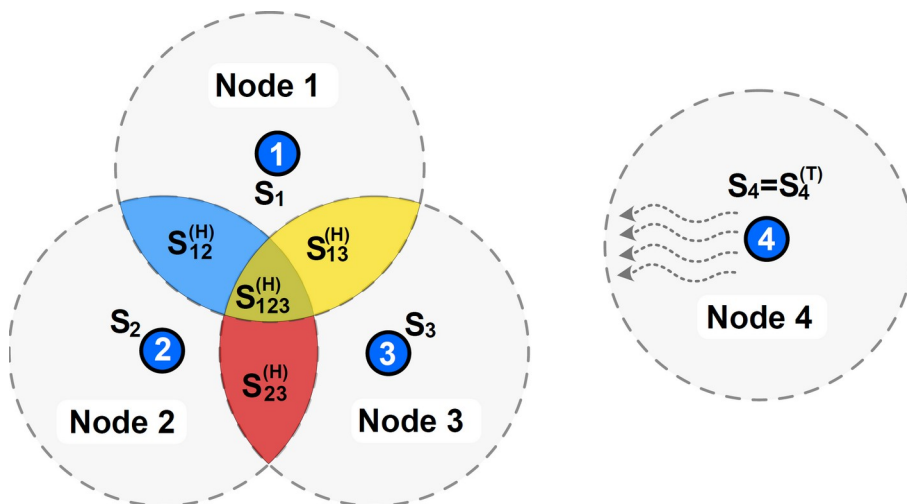
Sobol index:
Better for **node selection**



Shapley value:
Better for **contribution decomposition**

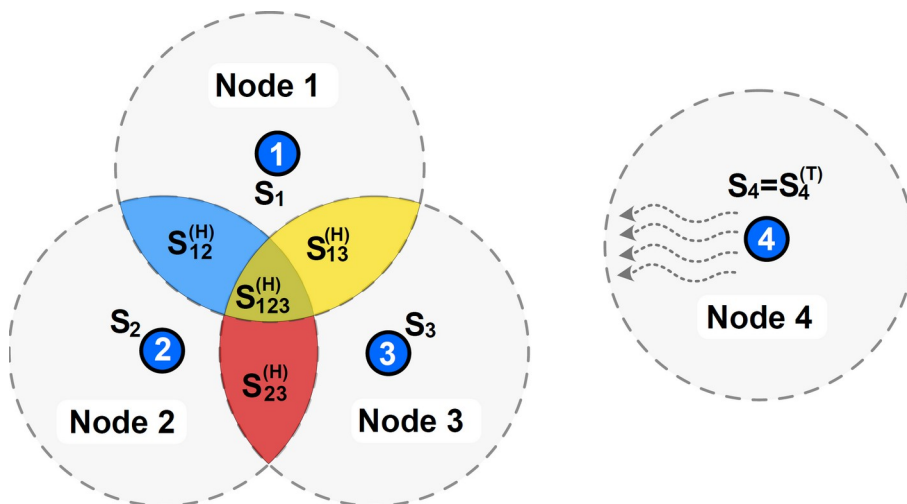
Sobol Indices

- First-order effect: $S_i := \frac{\mathbb{V}_{X_i} \mathbb{E}_{\mathbf{X}_{\sim i}} (Y \mid X_i)}{\mathbb{V}(Y)}$



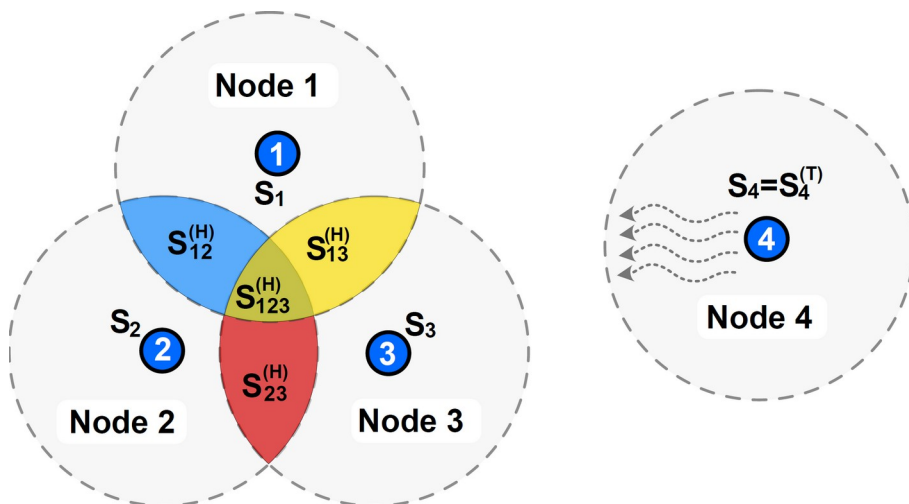
Sobol Indices

- For a variable set:
$$S_{ij} = \frac{\mathbb{V}_{X_i X_j} [\mathbb{E}_{\mathbf{x} \sim ij} (Y \mid X_i, X_j)]}{\mathbb{V}(Y)}$$



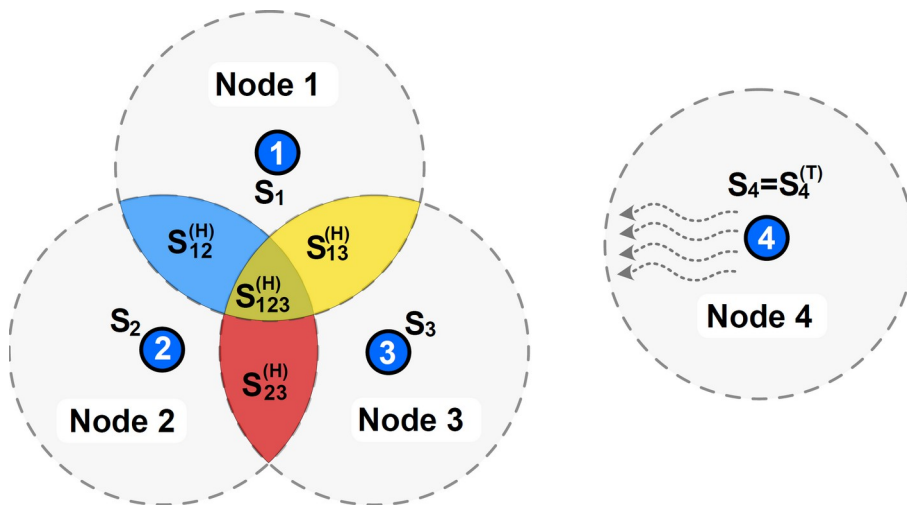
Sobol Indices

- Higher-order effect: $S_{ij}^{(H)} = \frac{\mathbb{V}_{ij}^{(H)}}{\mathbb{V}(Y)} = S_{ij} - S_i - S_j$
- We have:
$$\sum_i S_i + \sum_i \sum_{j>i} S_{ij}^{(H)} + \cdots + S_{12\dots n}^{(H)} = 1.$$



Sobol Indices

- Total effect: $S_i^{(T)} = \frac{\mathbb{E}_{\mathbf{X}_{\sim i}} (\mathbb{V}_{X_i} (Y \mid \mathbf{X}_{\sim i}))}{\mathbb{V}(Y)}$

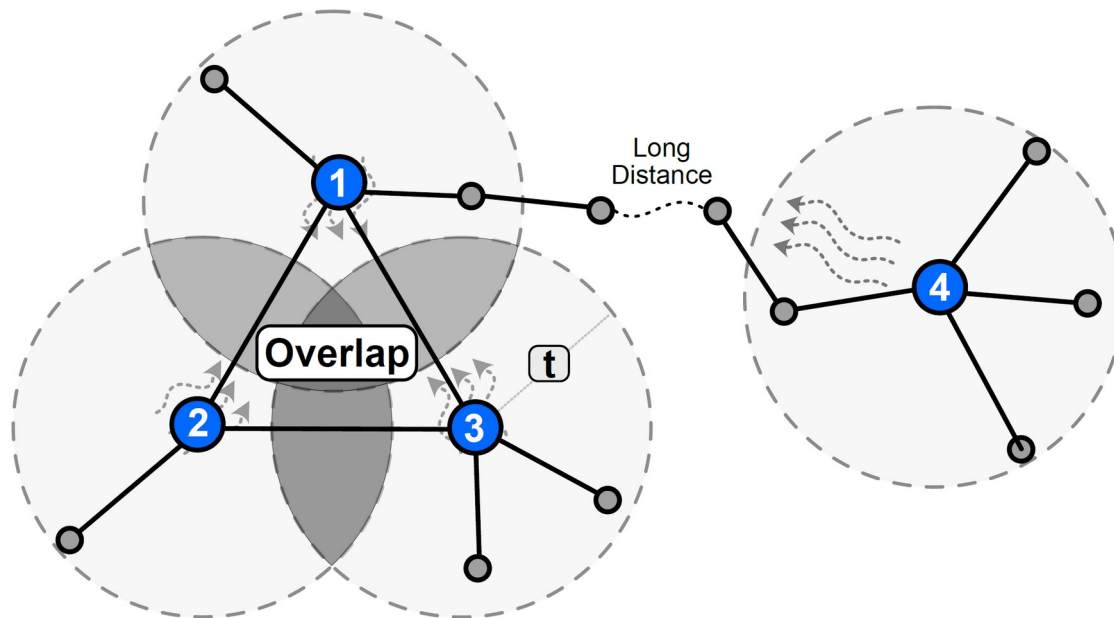


Agenda

- Motivation
- Introduction to **Sobol** Indices
- Influence Decomposition with **Sobol** Indices
- Proposed Method for IM
- Empirical Experiment

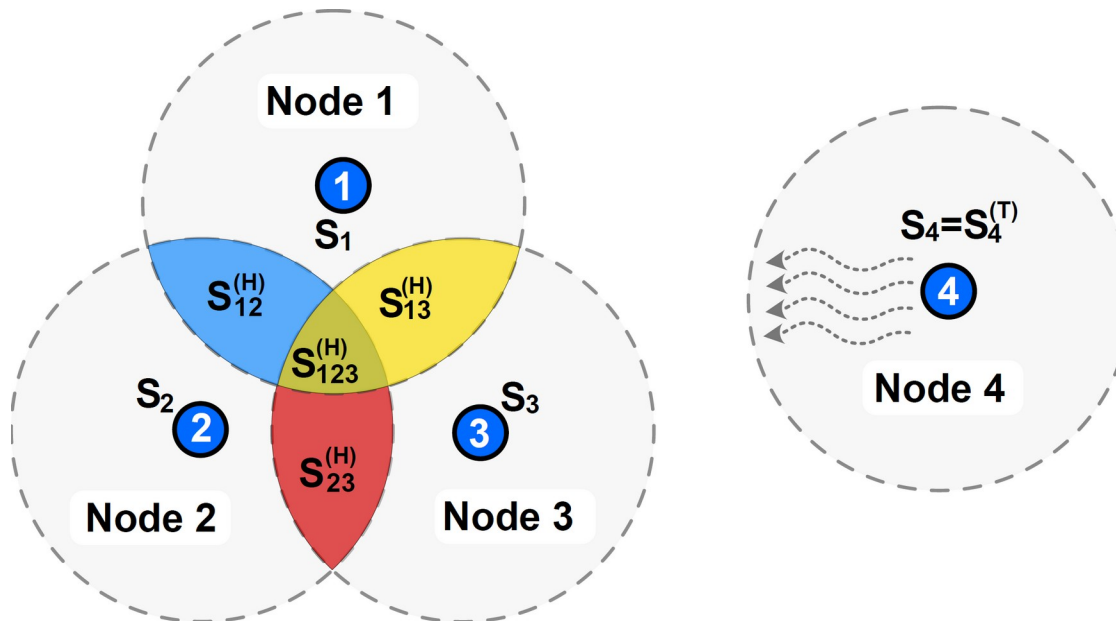
Explanatory Method

- Define node selection as **binary** variables.
- Variance from the **uncertainty** in the influence process.



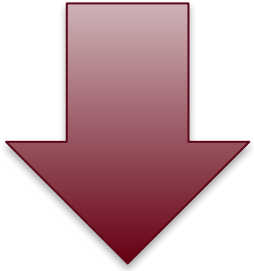
Explanatory Method

- Define node selection as **binary** variables.
- Variance from the **uncertainty** in the influence process.



Explanatory Method

- **Original first-order** Sobol index of **single variable:**



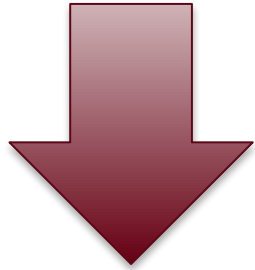
$$S_i := \frac{\mathbb{V}_{X_i} \mathbb{E}_{\mathbf{X}_{\sim i}} (Y \mid X_i)}{\mathbb{V}(Y)}$$

- **IM first-order** Sobol: contribution of **single seed:**

$$S_i = \frac{(\sum_{\Omega_{\sim i}} \Delta_{\Omega_{\sim i}})^2}{4^k \cdot \mathbb{V}(Y)} \quad \text{where} \quad \Delta_{\Omega_{\sim i}} := |Y_{\Omega_i=1} - Y_{\Omega_i=0}|$$

Explanatory Method

- **Original first-order** Sobol index for **a variable set:**



$$S_{ij} = \frac{\mathbb{V}_{X_i X_j} [\mathbb{E}_{\mathbf{X} \sim ij} (Y \mid X_i, X_j)]}{\mathbb{V}(Y)}$$

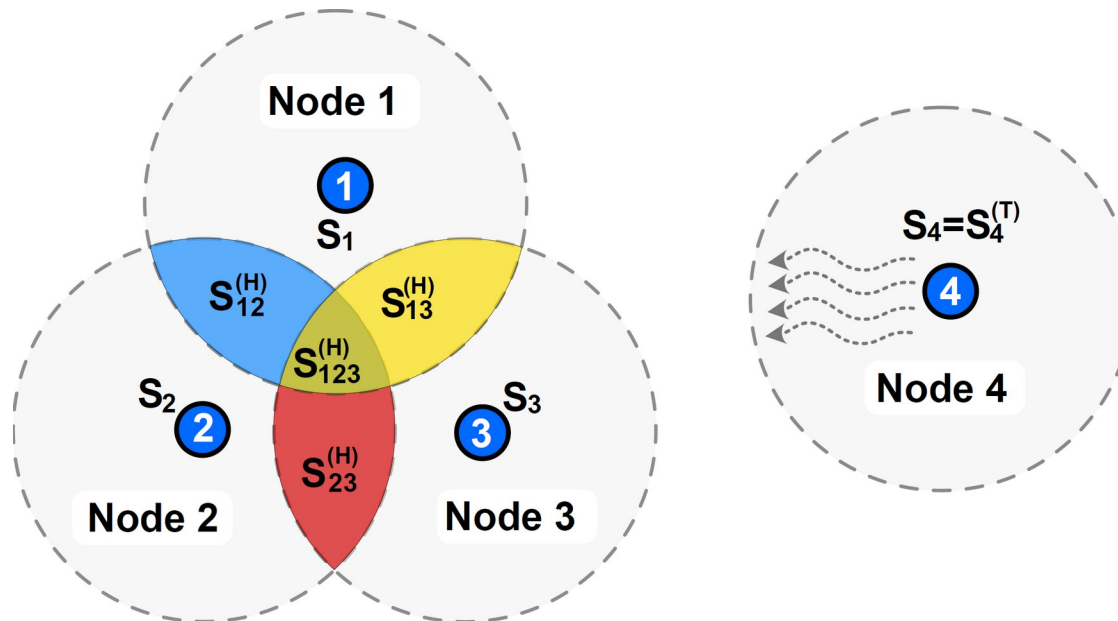
- **IM first-order** Sobol: pure contribution of **a seed set:**

$$S_{\Psi} = \frac{\sum_{\Psi} (\sum_{\Omega \sim \Psi} (Y \mid \Psi) \cdot 2^{s-k} - \mathbb{E}(Y))^2}{2^s \cdot \mathbb{V}(Y)}$$

Explanatory Method

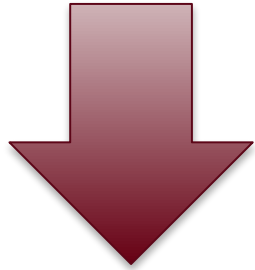
- Higher-order contribution of a seed:

$$S_{\Psi}^{(H)} = S_{\Psi} - \sum_{i=1}^s S_i - \sum_{|\zeta|=2}^{| \zeta |=(s-1)} \sum_{\zeta} S_{\zeta}^{(H)}$$



Explanatory Method

- **Original total effect Sobol of single variable:**



$$S_i^{(T)} = \frac{\mathbb{E}_{\mathbf{X}_{\sim i}} (\mathbb{V}_{X_i} (Y \mid \mathbf{X}_{\sim i}))}{\mathbb{V}(Y)}$$

- **IM total effect Sobol: global contribution of a seed:**

$$S_i^{(T)} = \frac{\sum_{\Omega_{\sim i}} (\Delta_{\Omega_{\sim i}})^2}{2^{k+1} \cdot \mathbb{V}(Y)}$$

Agenda

- Motivation
- Introduction to **Sobol** Indices
- Influence Decomposition with **Sobol** Indices
- **Proposed Method for IM**
- Empirical Experiment

Proposed Model: SIM

Algorithm 1 Sobol Influence Maximization (SIM)

Input: Graph $G = (\mathcal{V}, \mathcal{E}, A)$, budget constraint k , number of propagation steps t , IM algorithm M

Output: A k -sized seed set Ω , i.e., $|\Omega| = k$

```
1: /* Collecting */
2: Set a over-selection parameter  $a \in \mathbb{R}^+, a > 1$ 
3:  $\Omega \leftarrow M(G, \lceil ak \rceil, t)$ 
4: /* 2, Pruning */
5: while  $|\Omega| > k$  do
6:   for seed  $i$  in  $\Omega$  do
7:     Calculate  $S_i^{(T)}$ 
8:   end for
9:    $i \leftarrow \arg \min_i S_i^{(T)}$ 
10:   $\Omega \leftarrow \Omega_{\sim i}$ 
11: end while
```

Collecting:

$\lceil ak \rceil$ candidates are identified with SOTA proxy-based IM algorithms, where a is an arbitrary constant.

Proposed Model: SIM

Algorithm 1 Sobol Influence Maximization (SIM)

Input: Graph $G = (\mathcal{V}, \mathcal{E}, A)$, budget constraint k ,
number of propagation steps t , IM algorithm M

Output: A k -sized seed set Ω , i.e., $|\Omega| = k$

```
1: /* 1, Collecting */
2: Set a over-selection parameter  $a \in \mathbb{R}^+, a > 1$ 
3:  $\Omega \leftarrow M(G, \lceil ak \rceil, t)$ 
4: /* Pruning */
5: while  $|\Omega| > k$  do
6:   for seed  $i$  in  $\Omega$  do
7:     Calculate  $S_i^{(T)}$ 
8:   end for
9:    $i \leftarrow \arg \min_i S_i^{(T)}$ 
10:   $\Omega \leftarrow \Omega_{\sim i}$ 
11: end while
```

Pruning:

Extra nodes with lower Sobol total index need to be removed, one at a time.

Proposed Model: SIM

Algorithm 1 Sobol Influence Maximization (SIM)

Input: Graph $G = (\mathcal{V}, \mathcal{E}, A)$, budget constraint k ,
number of propagation steps t , IM algorithm M

Output: A k -sized seed set Ω , i.e., $|\Omega| = k$

```
1: /* 1, Collecting */
2: Set a over-selection parameter  $a \in \mathbb{R}^+, a > 1$ 
3:  $\Omega \leftarrow M(G, \lceil ak \rceil, t)$ 
4: /* Pruning */
5: while  $|\Omega| > k$  do
6:   for seed  $i$  in  $\Omega$  do
7:     Calculate  $S_i^{(T)}$ 
8:   end for
9:    $i \leftarrow \arg \min_i S_i^{(T)}$ 
10:   $\Omega \leftarrow \Omega_{\sim i}$ 
11: end while
```

Time complexity:

$$a * \mathcal{O}(M) + \mathcal{O}(2^{ak}r) \sim \mathcal{O}(M)$$

Agenda

- Motivation
- Introduction to **Sobol** Indices
- Influence Decomposition with **Sobol** Indices
- Proposed Method for IM
- Empirical Experiment

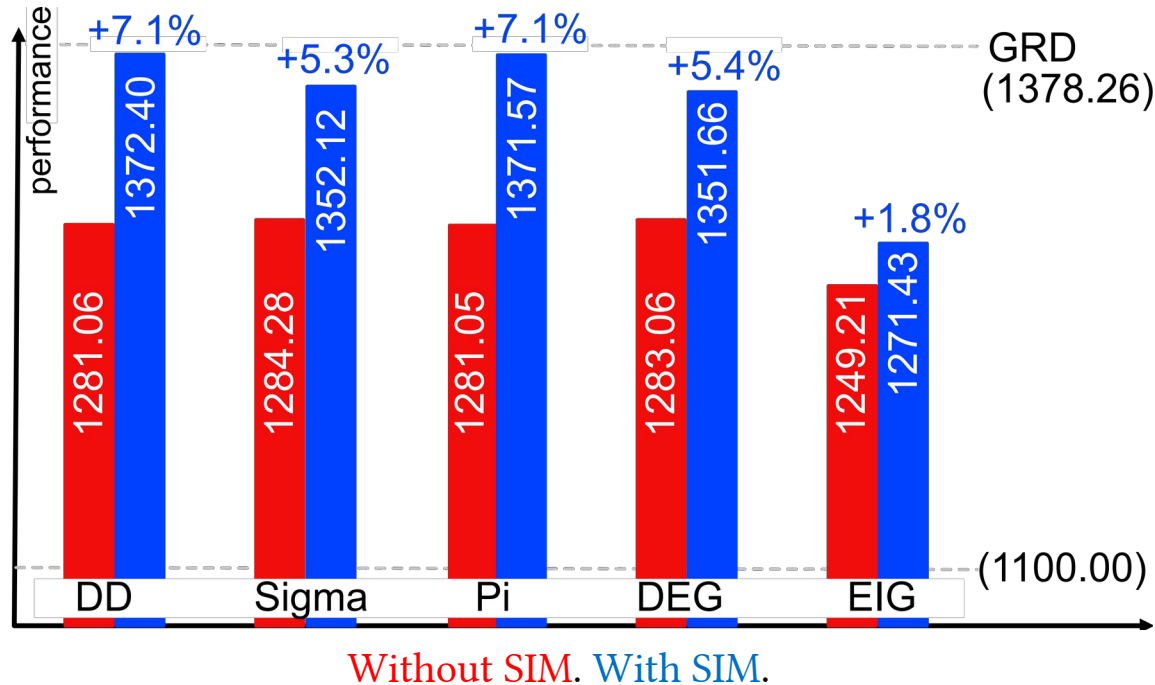
Experimental Setup

- Datasets:
 - Five real-world networks
 - Planetoid: Cora, CiteSeer, PubMed
 - Amazon Computer, Amazon Photo
 - Two synthetic networks
 - BA power-law graphs
 - ER graphs

Experimental Setup

- Six baselines:
 - **GRD** (oracle model)
 - simulation-based greedy algorithm (Kempe et al., 2003) on Cora only.
 - **Heuristics**
 - degree and Eigen-centrality on both LT and IC.
 - **Proxies** designed for IC:
 - Discounted Degree (Chen et al., 2009)
 - Sigma (Yan et al., 2019)
 - Pi (Zhang et al., 2022)

Results



Performances on Cora with IC comparing to GRD.

- **Effectiveness:**
 - SIM **boosts spread by 7%** over DD
 - SIM almost reaches the same level of GRD
- **Efficiency:**
 - GRD: 34 871.65 seconds
 - SIM+DD: 965.03 seconds, **accelerating by 36+ times**

Results

		CiteSeer (n=2120)	PubMed (n=19717)	Computers (n=13471)	Photo (n=7535)	ER (n=5000)	SW (n=5000)
DEG	W/O	618.90 $\pm$ 34.37	10948.96 $\pm$ 164.77	9221.05 $\pm$ 101.55	3523.68 $\pm$ 81.76	4608.97 $\pm$ 68.80	1284.44 $\pm$ 125.10
	W	698.20 $\pm$ 37.62	11258.95 $\pm$ 156.70	9254.33 $\pm$ 118.61	3581.44 $\pm$ 80.87	4619.67 $\pm$ 67.84	1386.88 $\pm$ 131.43
EIG	W/O	423.41 $\pm$ 20.34	5039.46 $\pm$ 411.31	9216.60 $\pm$ 105.91	3456.62 $\pm$ 77.64	4601.65 $\pm$ 69.25	1084.60 $\pm$ 115.95
	W	485.83 $\pm$ 24.79	7639.77 $\pm$ 439.76	9246.57 $\pm$ 98.83	3568.40 $\pm$ 79.86	4628.89 $\pm$ 63.24	1227.93 $\pm$ 126.65
DD	W/O	672.80 $\pm$ 37.99	11067.78 $\pm$ 172.69	9219.99 $\pm$ 94.94	3523.22 $\pm$ 78.76	4593.98 $\pm$ 72.22	1311.43 $\pm$ 121.18
	W	698.79 $\pm$ 36.01	11334.28 $\pm$ 178.05	9263.77 $\pm$ 99.24	3612.01 $\pm$ 87.54	4625.59 $\pm$ 65.79	1416.00 $\pm$ 117.92
Sigma	W/O	423.82 $\pm$ 20.89	11084.95 $\pm$ 162.37	9214.13 $\pm$ 100.97	3522.06 $\pm$ 83.07	4592.14 $\pm$ 67.82	1142.68 $\pm$ 105.22
	W	694.06 $\pm$ 36.18	11311.20 $\pm$ 144.72	9214.29 $\pm$ 98.72	3578.63 $\pm$ 82.87	4613.20 $\pm$ 67.19	1226.70 $\pm$ 113.71
Pi	W/O	619.03 $\pm$ 33.78	10960.65 $\pm$ 174.90	9212.89 $\pm$ 97.13	3524.33 $\pm$ 81.99	4597.28 $\pm$ 65.79	1163.93 $\pm$ 121.87
	W	697.63 $\pm$ 35.81	11345.50 $\pm$ 151.77	9251.23 $\pm$ 117.30	3583.76 $\pm$ 83.45	4624.04 $\pm$ 62.64	1355.83 $\pm$ 120.89

IM performance (i.e., influence spread) with (W) and without (W/O) SIM under IC model.

Combining the degree discount heuristic with SIM always achieves the best or the second-best performance among the algorithms.

Results

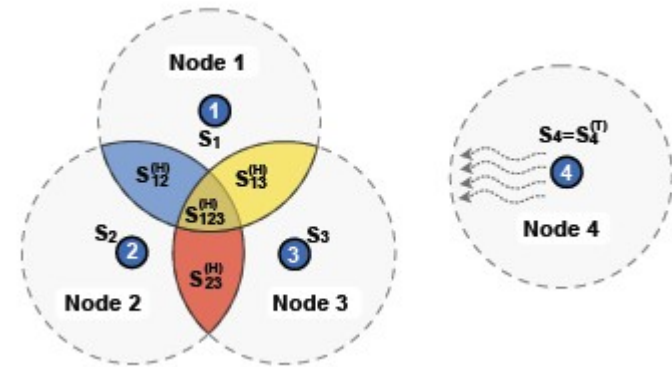
		CiteSeer (n=2120)	Cora (n=2485)	PubMed (n=19717)	Computers (n=13471)	Photo (n=7535)	ER (n=5000)	SW (n=5000)
DD	W/O	.02	0.02	0.18	0.77	0.33	0.07	0.07
	W	94.15	108.72	724.74	2097.04	984.28	294.66	269.31
Pi	W/O	1.80	2.43	147.03	69.05	22.29	9.73	9.87
	W	97.30	113.62	1020.89	2260.50	1025.19	314.57	289.20
Sigma	W/O	1.33	1.82	110.75	52.11	16.63	7.28	7.31
	W	96.30	112.62	937.42	2222.02	1007.60	310.80	285.01
DEG	W/O	0.02	0.02	0.19	0.78	0.33	0.08	0.13
	W	94.25	109.29	665.92	2140.26	985.84	296.98	268.18
EIG	W/O	0.10	0.14	1.55	4.25	2.03	0.56	0.65
	W	93.35	109.21	645.34	2143.48	989.88	298.04	269.19

- The time cost of pruning extra nodes does not scale with the graph size.
- SIM runs faster on more sparse graphs.

Results

Node Index	Sobol Total	Marginal Contribution
1422	0.305 ± 0.001	47.782 ± 5.960
582	0.462 ± 0.004	194.512 ± 4.495
1214	0.175 ± 0.002	6.916 ± 5.014
2782	0.182 ± 0.005	7.220 ± 5.751
1943	0.154 ± 0.001	5.354 ± 4.434

The Legitimacy of Each Seed's Relative Contribution.



Pairs (node index)	1422	1422	1422	1422	2782
	582	1214	2782	1943	1943
Second-order Sobol	0.00	591.60	619.46	518.00	340.83
Distance	6	2	2	2	2

Pairs (node index)	582	582	582	1214	1214
	1214	2782	1943	2782	1943
Second-order Sobol	0.01	0.00	0.08	424.81	329.21
Distance	7	7	7	1	2

The second-order Sobol indices vs. distance.

- Sobol total has a strict and positive correlation with marginal contribution.
- Pair of nodes that far away from each other has low interaction identified by the second-order Sobol effect.

“second-order models provide an excellent trade-off in practice”

Q&A

Thank you!

Code Here ↓



My Website ↓

