Understanding Influence Maximization via Higher-Order Decomposition

Zonghan Zhang Zhiqian Chen

Department of Computer Science and Engineering, Mississippi State University
zz239@msstate.edu, zchen@cse.msstate.edu

Abstract

Given its vast application on online social networks, Influence Maximization (IM) has garnered considerable attention over the last couple of decades. Due to the intricacy of IM, most current research concentrates on estimating the first-order contribution of the nodes to select a seed set, disregarding the higher-order interplay between different seeds. Consequently, the actual influence spread frequently deviates from expectations, and it remains unclear how the seed set quantitatively contributes to this deviation. To address this deficiency, this work dissects the influence exerted on individual seeds and their higher-order interactions utilizing the Sobol index, a variance-based sensitivity analysis. To adapt to IM contexts, seed selection is phrased as binary variables and split into distributions of varying orders. Based on our analysis with various Sobol indices, an IM algorithm dubbed SIM is proposed to improve the performance of current IM algorithms by over-selecting nodes followed by strategic pruning. A case study is carried out to demonstrate that the explanation of the impact effect can dependably identify the key higher-order interactions among seeds. SIM is empirically proved to be superior in effectiveness and competitive in efficiency by experiments on synthetic and real-world graphs.

1 Introduction

As online social networks have drawn and maintained hundreds of millions of users during the past few decades, influence maximization (IM) attracts great attention [8, 16]. It is widely applied to viral marketing [4], rumor control [11], social recommendation [32], and infectious disease containment [17]. IM is an NP-hard task in which a k-sized seed set is selected to maximize the number of influenced nodes which is also called influence spread [13]. The ineffectiveness or unreliability of IM methods could result in the loss of millions of dollars or even human lives, so their performance and dependability are of the utmost importance.

Current IM research suffers two major flaws. **(1) Lack of tools for influence spread decomposition.** Most current research focuses on identifying a set of seeds that maximizes the expected influence spread. Other than that, which seed or seeds account for the greatest contribution to the final influence has not been discussed. Although some of the heuristics select seeds according to transparent measures such as centrality metrics, the process through which the seed set results in the final influence spread is still hidden. Nonetheless, due to the high stake in the quality of the solution, this explanation of the selected seed set is necessary. Meanwhile, existing tools that can disentangle and quantitatively evaluate the contributions of the input variables in a nonlinear function have various drawbacks: *Local interpretation methods*, including LIME [20] and Grad-cam [26], focus only on local neighborhoods and disregard the global impact of the seeds and their higher-order interactions. *Marginal contribution* [3] does not separate the higher-order interaction effects from the main effect thus fails to quantify the interactions among seeds. *The Shapley value* [22] distributes the higher-order interaction among the seeds and tends to underestimate the contribution of the nodes with larger influence overlaps with others [18]. Beyond that, none handle higher-order interactions in the IM scenario well.

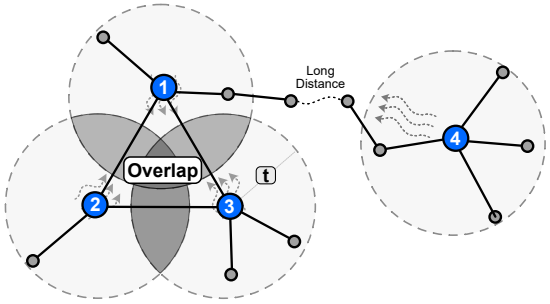

Figure 1: Interaction among the seeds.

(2) The imbalance of effectiveness and efficiency. Current seed selection methods are typically divided into two categories, simulation-based and proxy-based [16]. Unfortunately, neither can offer consistently trustworthy solutions in a timely manner. *The simulation-based methods* are theoretically guaranteed to have a high approximation performance by greedily adding the node

with the highest marginal spread in each step [13]. However, multiple rounds of simulations are required to evaluate the marginal spread for each candidate node in each round, making the heuristic extremely slow when the graph size is large [1]. For the sake of efficiency, *the proxy-based IM algorithms* emphasize the first-order contribution while disregarding the interaction efficacy among the seeds [5, 31, 33]. Nevertheless, due to the combinatorial impact in the seed set, the actual influence spread is heavily influenced by this interaction [13, 5, 16]. Thus, there is no assurance that the proxy-based algorithms can reach a good approximation ratio. Figure 1 illustrates an example of this interaction by depicting the influence distribution of a four-node seed set. Three of them (nodes 1, 2, and 3) are close to one another, whereas the fourth is far away. During t stages of the propagation, the influence flows propagated from the first three seeds may overlap (gray areas), canceling a portion of their overall influence spread and failing to maximize the joint coverage. In IM community, most research still focuses on searching for faster but less accurate simulation methods [6, 2] or designing better proxies that usually come with higher computational loads [31, 33]. Still, the gap between effectiveness and efficiency is yet to be filled.

To address these issues, we introduce a global sensitivity analysis method, namely the Sobol indices [27, 29], as an effective method to decompose a seed set's influence spread. In particular, the global contribution of a seed node is measured using the Sobol total index, which reflects its overall contribution to the influence spread. The contribution of the node is then explained using first- and higher-order Sobol indices. Notably, the interactions between seed nodes are identified quantitatively with a higher-order index, which explains why most first-order approaches fail to identify the contributions of nodes appropriately. The Sobol indices can overcome the shortcomings of the alternative tools mentioned above: **(1)** It is a global sensitivity method that considers a seed's global impact within the seed set. **(2)** The first-order effect and the higher-order interactions can be separated with a clear cut utilizing the corresponding Sobol indices. **(3)** A seed's Sobol total index includes the complete contribution of all of its higher-order interactions with the rest of the seed set. Hence it never overestimates the influence like Shapley value does [18].

After decomposing the influence spread towards each node, we rank the nodes according to their relative importance defined by their contributions. With this information, we propose a light-weighted method named SIM to find a balance between effectiveness and efficiency. Specifically, SIM selects excessive candidates and uses the Sobol indices to identify the ones that need to be excluded. Our primary contributions include:

- **Explain and decompose the influence spread**: Given any seed set and the corresponding graph, each seed's contribution to the influence spread can be measured with the proposed explanatory method regardless of the IM algorithm that generates the seed set. Thus, this method can serve as a universal post-hoc explanation of an IM algorithm. To the best of our knowledge, this paper serves as the first discussion on quantifying the influence overlap among the seeds.
- **Provide a new IM schema**: By combining the simulation with the proxies, SIM overcomes the scalability and dependability issues in the literature. It suggests a new direction for IM research and provides a new schema for designing IM methods.
- **Conduct extensive empirical experiments**: we prove that the Sobol indices can reliably evaluate the contribution of the nodes in the seed set with a case study. Real-world and synthetic datasets are employed to demonstrate that SIM achieves superior performance in a reasonable amount of time.

2 Related Work

Influence Maximization. Given that IM is NP-hard, researchers attempt to find a feasible solution with the best performance. As the first approximation attempt, a simulation-based greedy algorithm was proposed in [13], which is not scalable. Similarly, a thread of simulation-based methods is developed to increase the performance or reduce the complexity [15, 10]. Although great effort has been made to accelerate the process of the simulation-based methods, the complexity is still unacceptably high for the enormous online social networks [1, 28]. Most importantly, the opaqueness of the simulation makes this thread of methods impossible to be explained or improved by evaluating the diffusion process [16]. To alleviate the heavy computational burden of simulations, researchers turned to proxy-based methods in which the spreading power of the nodes is estimated by specific proxies. They started from simple heuristic measures such as degree, PageRank [19], eigen-centrality [34] and later turned to influence-aware or diffusion model-aware proxies [5, 14, 31, 33] to better estimate the influence spread brought by the seeds.

Variance-Based Sensitivity Analysis. Sensitive analysis studies the proportion based on which the uncertainty lying within the output of a model is distributed to the multiple sources of uncertainty in the input variables [25]. Variance-based methods take a major part in

sensitive analysis and have been dated back to 1970s [7]. Since then, these methods have been well studied and widely adopted by researchers and practitioners [24]. Among them, the Sobol indices [27] is considered a significant milestone. However, variance-based methods cannot be directly utilized on IM problems since the input variables are not clearly defined and the influence spread has not been represented as a nonlinear function. In the following section, we will adapt the Sobol indices and a few of its important inheritors to the IM context and utilize them to provide a human-understandable explanation for a given seed set.

3 Problem Setup

A network is denoted by a bi-directional graph $G = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ and $\mathcal{E}$ represent vertices and edges respectively. Given the graph G, a seed budget $k \in \mathbb{N}^+$, and an influence maximization method M, a k-size seed set $\Omega = M(G, k)$ can be generated to approximately maximize the expected influence spread.

(1) Our first objective is to develop ***an explanatory method*** that can explain the total contribution of a single candidate seed node from a higher-order perspective. The **T**otal contribution $f_i^{(T)}$ of seed i towards the expected influence spread $\sigma(\Omega)$ can be decomposed to different orders of i such that:

$$f_i^{(T)} = \sum_{\Psi \subseteq \Omega} f_\Psi, \tag{3.1}$$

where Ψ is all the possible subsets containing node i, and f_Ψ is the contribution of the interaction involving all nodes within a seed set Ψ. Suppose $\Omega = \{i, j, k\}$, then $\Psi \in \{\{i\}, \{i, j\}, \{i, k\}, \{j, k\}, \{i, j, k\}\}$. Note that $\Psi = \{i\}$ means the first-order contribution. f_Ψ is the contribution of the highest-order interaction in Ψ. For example, when $\Psi = \{i, j, k\}$, f_Ψ represent the interaction among i, j, k together, with no pairwise interaction such as the interaction between i and j.

(2) Our second objective is to design ***an efficient and effective IM schema*** to identify the most influential seed set (Ω^*) of size k in terms of higher-order interactions, such that:

$$\Omega^* = \underset{\Omega}{\operatorname{argmax}}\Big(\sum_i f_{\{\Omega_i\}} + \sum_i \sum_{j>i} f_{\{\Omega_i, \Omega_j\}} + \cdots + f_\Omega\Big) \tag{3.2}$$

in which Ω_i is the i-th node of Ω.

4 Preliminary: Functional Analysis of Variance

The objective of the functional analysis of variance (ANOVA) is to assess the significance of input to output based on the variance relationship between them. In this section, we will introduce ANOVA with *first-order effect*, *higher-order effects* and *total effect*, which is defined by Sobol indices [23] and represent significance from different perspectives. Specifically, variance-based functional decomposition is employed to analyze the impact of each variable. Given any model of the form $Y = f(X_1, X_2, \ldots X_n)$, with Y a scalar, the steps of a variance-based framework are as follows:

$$Y = f_0 + \overbrace{\sum_i f_i}^{\text{first-order}} + \overbrace{\sum_i \sum_{j>i} f_{ij} \quad + \cdots + \quad f_{12\ldots n}}^{\text{higher-order}}, \tag{4.3}$$

where

$$\begin{aligned} f_0 &= \mathbb{E}(Y), \\ f_i &= \mathbb{E}_{\mathbf{X}_{\sim i}}(Y \mid X_i) - \mathbb{E}(Y), \\ f_{ij} &= \mathbb{E}_{\mathbf{X}_{\sim ij}}(Y \mid X_i, X_j) - f_i - f_j - \mathbb{E}(Y), \end{aligned}$$

and similarly for the higher-order terms. X_i is the i-th variable, and $\mathbf{X}_{\sim i}$ denotes the variable set $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ that excludes X_i. $\mathbb{E}$ denotes the expectation. So f_i can be explained as the expected value change of Y taken over $\mathbf{X}_{\sim i}$ before and after X_i is fixed, deriving the contribution of X_i to Y. Similarly, f_{ij} means the expected value change of Y before and after fixing X_i and X_j. We will demonstrate below how to calculate the variance of a variable in various orders.
(1) First-Order Effect, denoted as $\mathbf{V}_i$, is the impact through one single variable only, which is represented by variance $\mathbb{V}$ accounted for X_i and it can be written as:

$$\mathbf{V}_i := \mathbb{V}(f_i) = \mathbb{V}_{X_i}[\overbrace{\mathbb{E}_{\mathbf{X}_{\sim i}}(Y \mid X_i)}^{\text{inner expectation}}], \tag{4.4}$$

since $\mathbb{V}_{X_i}[\mathbb{E}(Y)] = 0$. $\mathbb{V}_{X_i}(\cdot)$ represents the variance of argument $(\cdot)$ taken over X_i .The meaning of the *inner expectation* is that the mean of Y is taken over all possible values of $\mathbf{X}_{\sim i}$ while keeping X_i fixed. The outer variance ($\mathbb{V}_{X_i}$) is taken over all possible values of X_i. The associated first-order sensitivity is obtained by normalization:

$$S_i = \frac{\mathbf{V_i}}{\mathbb{V}(Y)}. \tag{4.5}$$

(2) Higher-Order Effect, using a second-order interaction effect between X_i and X_j as an example, can be expressed as:

$$\mathbf{V}_{ij}^{(H)} := \mathbb{V}[f_{ij}(X_i, X_j)] = \mathbf{V}_{ij} - \mathbf{V}_i - \mathbf{V}_j, \tag{4.6}$$

which means the variance brought by the interactions between variables i and j. $\mathbf{V}_{ij}$ means the first-order variance of the variable set $\{i, j\}$, and it considers the variance of $\{i\}, \{j\}$, or$\{i, j\}$. Extending Equation 4.4 from an individual variable to a variable set, we have:

$$\mathbf{V}_{ij} = \mathbb{V}_{X_i X_j}\left[\mathbb{E}_{\mathbf{X}_{\sim ij}}(Y \mid X_i, X_j)\right]. \tag{4.7}$$

Further, Equation 4.6 can be normalized by $\mathbb{V}(Y)$:

$$S_{ij}^{(H)} = \frac{\mathbf{V}_{ij}^{(H)}}{\mathbb{V}(Y)} = S_{ij} - S_i - S_j, \tag{4.8}$$

which means that the second-order effect is the first-order effect of the variable set minus the first-order effects of both individual variables. Similarly, we can calculate the third-order interaction among variables X_i, X_j, and X_h as

$$S_{hij}^{(H)} = S_{hij} - S_h - S_i - S_j - S_{hi}^{(H)} - S_{hj}^{(H)} - S_{ij}^{(H)}, \tag{4.9}$$

and so on. Given there are totally n variables in the model, we have

$$\sum_i S_i + \sum_i \sum_{j>i} S_{ij}^{(H)} + \cdots + S_{12\ldots n}^{(H)} = 1. \tag{4.10}$$

(3) Total Effect is the impact through one variable and all the interactions with the other variables, i.e., first- and higher-order effects. It can be calculated with [12]:

$$S_i^{(T)} = \frac{\mathbb{E}_{\mathbf{X}_{\sim i}}\left(\mathbb{V}_{X_i}\left(Y \mid \mathbf{X}_{\sim i}\right)\right)}{\mathbb{V}(Y)}, \tag{4.11}$$

where $S_i^{(T)}$ measures the total effect, i.e. sum of all first- and higher-order effects involving X_i.

5 Higher-Order Influence Decomposition

We will use the Sobol indices described in the preliminary to examine the current IM algorithms. Specifically, we will discuss how to (1) **explain** how higher-order effect characterizes the contribution of which first-order methods fail to do; and (2) **estimate** the contribution of node selection with total effect.

As discussed in the last section, sensitivity analysis is based on continuous variables of a nonlinear function. Nevertheless, the node selection problem in IM has not been described as such a function before. To bridge the gap, we define node selection as binary variables in the IM problem, representing whether a seed is included in the optimal seed set since a node can be either selected or not while it can not be partially selected. Specifically, a uniform distribution is employed such that each variable can be either 0 or 1 with a 50%/50% probability. Similar to Sobol notation, Ω_i denotes the selection decision for the i-th seed (1 for selected, 0 for not). This adaptation would greatly simplify the estimation of the Sobol indices since the variables in the model are bounded to be binary, limiting the number of combinations to 2^k. Comparatively, the time complexity of calculating the Sobol indices of a model including not only binary variables is very high. For discrete variables, the number of combinations increases exponentially along with the number of categories. And continuous variables can be considered as categorical variables that have infinite categories.

With the binary variables representing the selection decisions of the nodes, the randomness in the model is brought by the uncertainty in the influence process. In our research, we adopt both the IC model and the LT model [13] to mimic the influence process. The propagation spreads in multiple discrete time steps t_i. At t_0, some initial nodes are active as seeds. The uncertainty is brought by the propagation probability for the IC model and the random assignment of the node thresholds in each round of simulation for the LT model. Details are discussed in Appendix.

5.1 Explain with First- and Higher-order Effects.

To explain the significance of the seed, its influence spread is decomposed into first- and higher-order effects. We will illustrate how to calculate first-order and higher-order impacts in an IM setting and identify which higher-order relationship contributes so substantially that a first-order analysis alone is insufficient.

(1) First-Order Effect in IM. The first-order Sobol index of the node i evaluates the influence spread it brings with no overlaps with other nodes. This is the area activated by the node i as the seed, while no other seeds could successfully activate it. Representing the selection of the node i with a binary variable, we can calculate the first-order Sobol index inspired by Equation 4.5:

$$S_i = \frac{\mathbb{V}_{\Omega_i}(\mathbb{E}_{\Omega_{\sim i}}(Y|\Omega_i))}{\mathbb{V}(Y)} = \frac{1}{4^k \cdot \mathbb{V}(Y)}[\sum_{\Omega_{\sim i}} Y_{\Omega_i=1} - Y_{\Omega_i=0}]^2, \tag{5.12}$$

where $\Omega_{\sim i}$ represents a subset of the seed set that excludes the i-th seed. See Appendix for transformation details of Equation 5.12. Define $\Delta_{\Omega_{\sim i}} := |Y_{\Omega_i=1} - Y_{\Omega_i=0}|$ as the difference of influence spreads of Ω_i given seed sets $\Omega_{\sim i}$, we can rewrite Equations 5.12 as:

$$S_i = \frac{(\sum_{\Omega_{\sim i}} \Delta_{\Omega_{\sim i}})^2}{4^k \cdot \mathbb{V}(Y)}, \tag{5.13}$$

which will be used to quantify the first-order effect of certain nodes and to compare the higher-order effects.

(2) Higher-Order Effect in IM. In the IM scenario, the higher-order Sobol indices can quantify the influence overlaps among the nodes and help with identifying the critical overlaps that make the IM algorithms misjudge the value of selecting a certain node as one of the seeds. The larger a node's total higher-order Sobol indices are, the more the actual influence spread it results in would shrink from its value identified by the IM algorithms. These indices can be calculated either by summing the higher-order Sobol indices from all orders together or by subtracting the first-order Sobol index from the Sobol

total index, The calculation of the higher-order Sobol indices can also be adapted to the IM settings and simplified. Based on 4.4, the first-order Sobol index of a size-s subset $\Psi \subseteq \Omega$ can be calculated by:

(5.14)
$$S_\Psi = \frac{\mathbb{V}_\Psi(\mathbb{E}_{\Omega_{\sim\Psi}}(Y|\Psi))}{\mathbb{V}(Y)} = \frac{\sum_\Psi(\sum_{\Omega_{\sim\Psi}}(Y|\Psi)\cdot 2^{s-k} - \mathbb{E}(Y))^2}{2^s \cdot \mathbb{V}(Y)}.$$

See Appendix for transformation details of Equation 5.14. The higher-order Sobol indices can be calculated in an iterative manner, starting from the second order to the k-th order. Specifically, the s-th order Sobol indices of Ψ whose $|\Psi| = s$ is:

$$S_\Psi^{(H)} = S_\Psi - \sum_{i=1}^{s} S_i - \sum_{|\zeta|=2}^{|\zeta|=(s-1)} \sum_{\zeta} S_\zeta^{(H)}, \tag{5.15}$$

where $\zeta \subseteq \Psi$. Note that there exist multiple subsets for each order. For example, suppose $\Psi = \{1,2,3\}$ as in Figure 2, the implementation of Equation 5.15 is:

$$S_{123}^{(H)} = S_{123} - S_1 - S_2 - S_3 - S_{12}^{(H)} - S_{23}^{(H)} - S_{13}^{(H)} \tag{5.16}$$

Similarly, it is possible to compute the fourth-order and even higher-order effects. Intuitively, though, higher-order relationships have fewer effects on average, such as $S_{123}^{(H)}$ being smaller than $S_{12}^{(H)}$, yet incur exponentially rising computing costs. In light of this, this study will investigate the second-order contributions to influence propagation, which, as proved by our experiments, are sufficient to explain why first-order is insufficient. Various higher-order selections can be expanded as necessary.

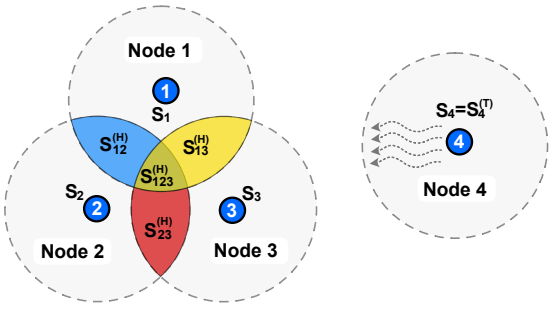

Figure 2: Sobol-based Influence decomposition.

5.2 Estimate influence spread of Node Set with Total Effect. Given a seed set Ω and the graph G representing the network where the propagation happens, we can estimate the influence spread within t time steps or until the propagation fully stops. Intuitively, we can write the relationship between the influence spread and the seed set as:

$$g : \Omega \to \sigma, \tag{5.17}$$

where g is a function of seed set Ω whose size is smaller than or equal to k, and g predicts the influence spread σ based on the seed set Ω. To evaluate the contribution of a single seed within the set, we need to quantify the variance brought by including the seed in the seed set.

To facilitate the evaluation, an analysis of the total effect will be applied to node set. As discussed, total effect measures the amount of influence variance lost if node i is not included in the seed set. It includes not only the part of the network activated by node i, but also the part that could be activated by this node and some other nodes at the same time. Following Equation 4.11, the Sobol total index for a specific node in IM can be written as:

(5.18)
$$S_i^{(T)} = \frac{\mathbb{E}_{\Omega_{\sim i}}(\mathbb{V}_{\Omega_i}(Y \mid \Omega_{\sim i}))}{\mathbb{V}(Y)} = \frac{\sum_{\Omega_{\sim i}}(Y_{\Omega_i=1} - Y_{\Omega_i=0})^2}{2^{k+1}\cdot\mathbb{V}(Y)}.$$

Detailed transformation is shown in Appendix. Similar to Equation 5.13, Equations 5.18 can be rewritten as:

$$S_i^{(T)} = \frac{\sum_{\Omega_{\sim i}}(\Delta_{\Omega_{\sim i}})^2}{2^{k+1}\cdot\mathbb{V}(Y)}, \tag{5.19}$$

This will be used to evaluate a specific node's global contribution to IM. Note that the numerator in Equation 5.19 is the sum of the squares while the numerator in Equation 5.13 is the square of the sum.

5.3 SIM: The Proposed Algorithm. Now that the Sobol total index can be used to evaluate the significance of a seed within the seed set, we can deduce that deleting the node with the lowest Sobol total index will result in the most negligible influence loss when the budget restriction is reduced from k to $k-1$. Based on this finding, we propose a light-weighted plug-in approach named SIM as described in Algorithm 1. It can improve the performance of the mainstream IM method in two steps, namely **collecting** of seed candidates and **pruning** of the nodes with less impact on the influence spread while maintaining a good balance between effectiveness and efficiency.

(1) Collecting (line 1-3). In the first step of SIM, $\lceil ak \rceil$ candidates are selected with state-of-the-art IM algorithms (line 3), leading to extra $\lceil ak \rceil - k$ nodes beyond the budget. IM algorithms would select overall good seed nodes ignoring higher-order interaction among them, but there may be over- or under-estimated effects, which need to be corrected.

(2) Pruning (line 4-11). To satisfy the budget constraint, $\lceil (a-1)k \rceil$ seeds need to be removed from the $\lceil ak \rceil$ candidates. This selection process is accomplished by the Sobol total indices (line 7). Note that the Sobol total index is to measure the variance lost when holding one variable stable with all other variables still in the

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: /* *2, Pruning* */
5: **while** $|\Omega| > k$ **do**
6: **for** seed i in Ω **do**
7: Calculate $S_i^{(T)}$ (with Equation 5.19)
8: **end for**
9: $i \leftarrow \arg\min_i S_i^{(T)}$
10: $\Omega \leftarrow \Omega_{\sim i}$
11: **end while**

model. Thus, removing the node with the lowest index results in the least influence loss. The pruning of the extra nodes must be done one at a time iteratively until exactly k nodes are in the set. The Sobol total index for each seed left would be updated after each iteration, and the rank of the seeds might change. Thus, selecting the nodes with the largest value of indices from the candidates or removing the $\lceil ak \rceil - k$ lowest-ranking nodes at one time does not guarantee the best result.

Time Complexity Analysis. For heuristic algorithms with no updates to the proxies during the selection procedure, such as degree and Eigen-centrality, the time cost of selecting ak nodes is the same as selecting k. For those who update their proxies, the time for each node selection and proxy update iteration remains the same while the number of iterations rises from k to ak. Thus, the time complexity of the Collecting process is at most $a * \mathcal{O}(M)$ where $\mathcal{O}(M)$ is the time complexity of M. This complexity is equal to $\mathcal{O}(M)$ when we consider a as a constant. Assuming that it takes r rounds of simulations to estimate a seed's Sobol total index (line 7 in Algorithm 1), the total number of simulation rounds required for calculating the indices throughout the pruning process is $2^{ak} \cdot r + 2^{ak-1} \cdot r + \cdots + 2^{k+1} \cdot r = (2^{ak+k} - 2^{k+1}) \cdot r = \mathcal{O}(2^{ak} r)$. The time complexity concerning ak seems exponentially huge. However, a is a small constant, and k is the budget restriction that does not grow with the graph size. Besides, r is a linear factor that will not expand as the size of the problem grows and will be set to an acceptable value. Therefore the extra time cost is constant regardless of the graph size as long as the budget restriction stays the same. The overall time complexity of the Algorithm 1 is $a*\mathcal{O}(M)+\mathcal{O}(2^{ak} r) \sim \mathcal{O}(M)$ unless the budget constraint grows with the graph size.

6 Experiment

Synthetic and real-world datasets are used to test the proposed explanatory method and SIM. Code for the experiments is available at `https://github.com/oates9895/SIM`

6.1 Configurations. All experiments are carried out on a server with 32 AMD EPYC 7302P 16-Core processors and 32GB RAM. Simulations are performed by NDLib [21] which is an open-source tool for investigating diffusion processes and dynamics.

Datasets. The framework is tested on seven datasets. **(1)** Five real-world datasets[1] including *Cora, CiteSeer, PubMed, Amazon Computers, and Amazon Photo* are employed to mimic the sophisticated online social network structure. Since the IM problems traditionally focus on connected networks, the largest connected component of each graph is utilized as the network from which the seed set is selected. For Cora, CiteSeer, and PubMed, the edges are uniformly and randomly weighted between 0.40 to 0.80 for the IC model to reflect the varied activation probabilities between nodes. For Amazon Computers and Amazon Photo graphs, the edges are between 0.05 to 0.20 since those two graphs are denser, and the propagation can pass on with lower activation probabilities. For the LT model, the nodes are randomly assigned thresholds uniformly distributed between 0.01 to 0.20 in each round of simulation for each graph. **(2)** Two synthetic graphs representing pseudo social networks are generated using NetworkX [2]. The graphs include *connected Watts-Strogatz small-world graphs* (SW) [30] and *Erdős–Rényi random graphs* (ER) [9]. Each graph has 5000 nodes, and the average degree is approximately 10. The edges and nodes are weighted in a similar manner to the CiteSeer graph.

Baselines. To evaluate the legitimacy of our evaluation of the selected seed set and the corresponding IM method, we select a few popular IM algorithms as our candidates. SIM is compared with these baselines to evaluate its performance: **(1)** *Degree Centrality (DEG)*: The first k nodes with the highest degree centrality from the target graph are selected. Before each iteration, the degrees are updated such that the selected seeds are removed from the original graph. This heuristic is also known as SingleDiscount [5]. **(2)** *Eigenvector Centrality (EIG)*: Similarly, the first k nodes are selected based on their eigenvector centrality. **(3)** *Simulation-based greedy algorithm (GRD)* [13]: In each iteration, the node with the highest marginal influence spread, which is measured by the average of 1000 rounds of simulations, is added to

[1] Datasets at https://pytorch-geometric.readthedocs.io
[2] https://networkx.org

the seed set. **(4)** *Degree discount (DD)* [5]: The degree of each candidate node is discounted according to the likelihood of it being activated by the nodes that have been already selected. **(5)** *Sigma* [31]: The spreading powers of the nodes are estimated by $\sum_t I \cdot A^t$ where I is a unit column vector. **(6)** *Pi* [33]: The nodes spreading powers are estimated by $I \cdot (J - \prod_{r=1}(1 - A^r))$ where J is an all-one matrix and $\prod$ is the element-wise product of matrices. Among the six, GRD is run on Cora graph only with both the IC model and the LT model to show its lack of scalability, DEG and EIG are applied to both the IC model and the LT model, while DD, Sigma, and Pi are only applied to the IC model.

6.2 Results. The empirical study generally consists of three major parts: (1) **case study**: illustrate the effectiveness of the node contribution decomposition with a 5-node seed scenario. (2) **effectiveness verification**: demonstrate the effectiveness of the proposed algorithm against current IM algorithms. (3) **runtime analysis**: compare the SIM with the baseline IM methods to evaluate its time efficiency.

6.2.1 Case Study. To determine if the Sobol indices are capable of decomposing influence, we conduct a case study using a seed set created by DEG on the most linked component of CiteSeer as an illustration. Five seeds are chosen based on their degree centralities, and the IC model is utilized to estimate the Sobol indices. DEG is chosen as the sampling technique due to its simplicity and readability. And CiteSeer is chosen for computational efficiency because it is the smallest of the five graphs. The contribution of each seed inside the set is measured by the difference between the estimated influence spreads of the node before and after its inclusion in the seed set (represented as "marginal contribution"). The total Sobol indices for the five seeds are shown in Table 1. The substantial positive association between a node's Sobol total index and its marginal contribution demonstrates that this index is a reliable indicator of the node's contribution. In addition, we analyzed the overlap amount and the distance between the nodes. When all the nodes involved are closer to one another, the influence overlaps tend to be larger because the influence flows emanating from the seeds converge early in the propagation process. If at least one set of nodes are separated by a great distance, the overlap will be small. As shown in Table 2, the second-order Sobol indices are proportional to the distance between any two nodes.

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

Table 1: The Legitimacy of Each Seed's Relative Contribution.

Pairs (node index)	1422 582	1422 1214	1422 2782	1422 1943	2782 1943
Second-order Sobol	0.00	591.60	619.46	518.00	340.83
Distance	6	2	2	2	2
Pairs (node index)	582 1214	582 2782	582 1943	1214 2782	1214 1943
Second-order Sobol	0.01	0.00	0.08	424.81	329.21
Distance	7	7	7	1	2

Table 2: The second-order Sobol indices vs. distance.

6.2.2 Effectiveness Verification. Under the IC model, it takes the simulation-based greedy algorithm (GRD) 17758.43 seconds to find a 5-seed set on the largest connected component of the Cora graph, who has 2485 nodes. While the running times of SIM and the other baselines are at most 113.62 seconds, we could observe that the expected influence spreads of the generated seed sets are almost the same. The IM performances of the six heuristics with and without SIM are compared with GRD in Figure 3. It shows that after combined with SIM, the heuristics could achieve a performance almost as good as GRD. The experiment under the LT models shows a similar result. DEG with SIM achieves an influence spread of 1936.11 within 965.03 seconds while a similar performance of 1949.15 takes GRD 34871.65 seconds.

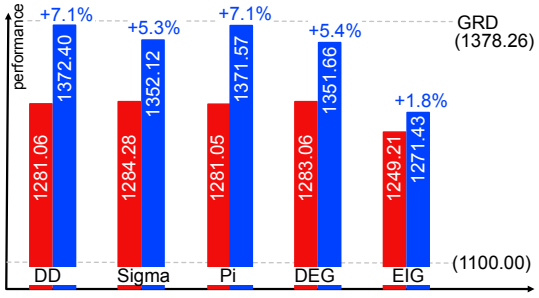

Figure 3: Performances on Cora with IC comparing to GRD. Red: Without SIM. Blue: with SIM.

Figure 3 also demonstrates that applying SIM enhances the IM performance with a ratio up to 7%. On the other six graphs, we compare the IM performance of the five baseline IM methods before and after they are combined with SIM. Without SIM, each IM algorithm generates a seed set consisting of 5 nodes. With SIM,

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

Table 3: IM performance (i.e., influence spread) with (W) and without (W/O) SIM under IC model. ***Italic bold*** for the best performance, **bold** for the second best.

each IM algorithm selects 10 candidate nodes during the Collecting procedure and then prunes 5 of them to meet the budget constraint. The final influence spread is measured by the mean and standard deviation of 1000 simulations to eliminate the uncertainty. The detailed result under the IC model is presented in Table 3. See Appendix for the result under the LT model.

We observe a significant increase on influence spread after SIM is applied in most scenarios, with a ratio up to 60%. The performance enhancement effect is greater when the original IM algorithm fails to generate an effective seed set. In most cases, the framework can achieve a boost of around 5% on the influence spread. Another significant finding is that combining the degree discount heuristic with SIM always achieves the best or the second best performance among the algorithms.

6.2.3 Runtime Analysis. It is also proved empirically that the time cost of generating the seed set increases as predicted. We observe that the increase ratios are smaller on larger graphs. This is because the time cost of pruning extra nodes is not related to the graph size, while the time cost of collecting candidate nodes scales with the graph size. Also, SIM runs faster on more sparse graphs since each round of simulations during the Sobol total indices estimation is faster. The details are demonstrated in Table 5 in Appendix.

7 Conclusion

This study presents a way of universally explaining the global effect of a seed set under the IM issue. Sobol's total index is used to assess the contribution of the seeds. The influence overlaps among seeds are quantified for the first time as an explanation using higher-order Sobol indices, which provide insight into the interaction effect and also demonstrate why the first-order method fails to accurately locate the ideal node set. The Sobol indices are calculated within the context of IM. Experiments indicate that the explanation we offered is applicable to graphs and IM algorithms. A revolutionary framework, SIM, is provided for improving the performance of any proxy-based IM algorithm. Experiments conducted on synthetic and real-world datasets demonstrated that our algorithm offers greater performance at an acceptable cost in terms of time.

Acknowledgement

This work was funded by the NSF IIS award # 2153369.

References

[1] A. Arora, S. Galhotra, and S. Ranu, *Debunking the myths of influence maximization: An in-depth benchmarking study*, in Proceedings of the 2017 ACM international conference on management of data, 2017, pp. 651–666.

[2] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, *Maximizing social influence in nearly optimal time*, in Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms, SIAM, 2014, pp. 946–957.

[3] A. Catav, B. Fu, Y. Zoabi, A. L. W. Meilik, N. Shomron, J. Ernst, S. Sankararaman, and R. Gilad-Bachrach, *Marginal contribution feature importance-an axiomatic approach for explaining data*, in International Conference on Machine Learning, PMLR, 2021, pp. 1324–1335.

[4] W. Chen, C. Wang, and Y. Wang, *Scalable influence maximization for prevalent viral marketing in large-scale social networks*, in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 1029–1038.

[5] W. Chen, Y. Wang, and S. Yang, *Efficient influence maximization in social networks*, in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 199–208.

[6] S. Cheng, H. Shen, J. Huang, G. Zhang, and X. Cheng, *Staticgreedy: solving the scalability-accuracy dilemma in influence maximization*, in Proceedings of the 22nd ACM international conference on Information & Knowledge Management, 2013, pp. 509–518.

[7] R. Cukier, C. Fortuin, K. E. Shuler, A. Petschek, and J. H. Schaibly, *Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. i theory*, The Journal of chemical physics, 59 (1973), pp. 3873–3878.

[8] P. Domingos and M. Richardson, *Mining the network value of customers*, in Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 2001, pp. 57–66.

[9] E. N. Gilbert, *Random graphs*, The Annals of Mathematical Statistics, 30 (1959), pp. 1141–1144.

[10] A. Goyal, W. Lu, and L. V. Lakshmanan, *Celf++ optimizing the greedy algorithm for influence maximization in social networks*, in Proceedings of the 20th international conference companion on World wide web, 2011, pp. 47–48.

[11] X. He, G. Song, W. Chen, and Q. Jiang, *Influence blocking maximization in social networks under the competitive linear threshold model*, in Proceedings of the 2012 siam international conference on data mining, SIAM, 2012, pp. 463–474.

[12] T. Homma and A. Saltelli, *Importance measures in global sensitivity analysis of nonlinear models*, Reliability Engineering & System Safety, 52 (1996), pp. 1–17.

[13] D. Kempe, J. Kleinberg, and É. Tardos, *Maximizing the spread of influence through a social network*, in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003, pp. 137–146.

[14] M. Kimura, K. Saito, and H. Motoda, *Blocking links to minimize contamination spread in a social network*, ACM Transactions on Knowledge Discovery from Data (TKDD), 3 (2009), pp. 1–23.

[15] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, *Cost-effective outbreak detection in networks*, in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007, pp. 420–429.

[16] Y. Li, J. Fan, Y. Wang, and K.-L. Tan, *Influence maximization on social graphs: A survey*, IEEE Transactions on Knowledge and Data Engineering, 30 (2018), pp. 1852–1872.

[17] M. E. Newman, *Spread of epidemic disease on networks*, Physical review E, 66 (2002), p. 016128.

[18] A. B. Owen, *Sobol'indices and shapley value*, SIAM/ASA Journal on Uncertainty Quantification, 2 (2014), pp. 245–251.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd, *The pagerank citation ranking: Bringing order to the web.*, tech. report, Stanford InfoLab, 1999.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, *" why should i trust you?" explaining the predictions of any classifier*, in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[21] G. Rossetti, L. Milli, S. Rinzivillo, A. Sirbu, D. Pedreschi, and F. Giannotti, *Ndlib: Studying network diffusion dynamics*, in 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2017, pp. 155–164.

[22] A. E. Roth, *The Shapley value: essays in honor of Lloyd S. Shapley*, Cambridge University Press, 1988.

[23] A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola, *Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index*, Computer Physics Communications, 181 (2010), pp. 259–270.

[24] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, *Global sensitivity analysis: the primer*, John Wiley & Sons, 2008.

[25] A. Saltelli and I. M. Sobol, *About the use of rank transformation in sensitivity analysis of model output*, Reliability Engineering & System Safety, 50 (1995), pp. 225–239.

[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, *Grad-cam: Visual explanations from deep networks via gradient-based localization*, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[27] I. M. Sobol, *Sensitivity analysis for non-linear mathematical models*, Mathematical modelling and computational experiment, 1 (1993), pp. 407–414.

[28] Y. Tang, X. Xiao, and Y. Shi, *Influence maximization: Near-optimal time complexity meets practical efficiency*, in Proceedings of the 2014 ACM SIGMOD international conference on Management of data, 2014, pp. 75–86.

[29] T. Wagener and F. Pianosi, *What has global sensitivity analysis ever done for us? a systematic review to support scientific advancement and to inform policy-making in earth system modelling*, Earth-science reviews, 194 (2019), pp. 1–18.

[30] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world'networks*, nature, 393 (1998), pp. 440–442.

[31] R. Yan, D. Li, W. Wu, D.-Z. Du, and Y. Wang, *Minimizing influence of rumors by blockers on social networks: algorithms and analysis*, IEEE Transactions on Network Science and Engineering, 7 (2019), pp. 1067–1078.

[32] M. Ye, X. Liu, and W.-C. Lee, *Exploring social influence for recommendation: a generative model approach*, in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012, pp. 671–680.

[33] Z. Zhang, S. Biswas, F. Chen, K. Fu, T. Ji, C.-T. Lu, N. Ramakrishnan, and Z. Chen, *Blocking influence at collective level with hard constraints (student abstract)*, (2022).

[34] L.-F. Zhong, M.-S. Shang, X.-L. Chen, and S.-M. Cai, *Identifying the influential nodes via eigen-centrality from the differences and similarities of structure*, Physica A: Statistical Mechanics and its Applications, 510 (2018), pp. 77–82.