

Beyond Worst-Case (In)approximability of Nonsubmodular Influence Maximization

Grant Schoenebeck Biaoshuai Tao *

Abstract

We consider the problem of maximizing the spread of influence in a social network by choosing a fixed number of initial seeds, formally referred to as the *influence maximization problem*. It admits a $(1 - 1/e)$ -factor approximation algorithm if the influence function is *submodular*. Otherwise, in the worst case, the problem is NP-hard to approximate to within a factor of $N^{1-\varepsilon}$. This paper studies whether this worst-case hardness result can be circumvented by making assumptions about either the underlying network topology or the cascade model.

First, we present strong inapproximability results for a very restricted class of networks called the *(stochastic) hierarchical blockmodel*, a special case of the well-studied *(stochastic) blockmodel* in which relationships between blocks admit a tree structure. We also provide a dynamic-program based polynomial time algorithm which optimally computes a directed variant of the influence maximization problem on hierarchical blockmodel networks. Our algorithm indicates that the inapproximability result is due to the bidirectionality of influence between agent-blocks.

Second, we present strong inapproximability results for a class of influence functions that are “almost” submodular, called *2-submodular*. Our inapproximability results hold even for any 2-submodular f fixed in advance. This model is motivated by many real life social network cascades. This result also indicates that the “threshold” between submodularity and nonsubmodularity is sharp, regarding the approximability of influence maximization.

1 Introduction

A *cascade* is a fundamental social network process in which a number of nodes, or agents, start with some property that they then may spread to neighbors. The importance of network structure on cascades has been shown to be relevant in a wide array of environments, including the adoption of products [5, 8, 18, 30], farming technology [15], medical practices [14], participation in microfinancing [4], and the spread of information over social networks [26].

A natural question, known as *the influence maximization problem* (INFMAX), is how to place a limited number k of initial seeds, in order to maximize the spread of the resulting cascade [17, 34, 24, 25, 32]. In order to study influence maximization, we first need to understand how cascades spread. Many cascade models have been proposed [2, 31, 38], and two simple examples are the Independent Cascade model [24, 25, 32] and the Threshold model [20]. In the Independent Cascade model, each newly infected node infects each currently uninfected neighbor in the subsequent round with some fixed probability p . In the Threshold model each node has a threshold (0, 1, 2, etc.) and becomes infected when the number of infected neighbors meets or surpasses that threshold.

In general, it is NP-hard even to approximate INFMAX to within $N^{1-\epsilon}$ of the optimal expected number of infections [25]. However, assuming that we are looking particular class of cascades, called

*Department of Computer Science, University of Michigan, Ann Arbor

submodular cascades, a straightforward greedy algorithm can efficiently find an answer that is at least a $(1 - 1/e)$ fraction of the optimal answer.

In submodular cascade models, such as the Independent Cascade model, a vertex’s marginal probability of becoming infected after a new neighbor is infected decreases with the number of previously infected neighbors [24]. Submodular cascade models are fairly well understood theoretically, and properties of these cascades are usually closely related to a network’s degree distribution and conductance [23]. Unfortunately, empirical research shows that many cascades are not submodular [35, 3, 27].

Cascade models that violate the submodularity property are called *nonsubmodular* cascades (or sometimes complex cascades). In nonsubmodular contagion models, like the Threshold model, the marginal probability of being infected may increase as more neighbors are infected. For example, if a vertex has a threshold of 2, then the first infected neighbor has zero marginal impact, but the second infected neighbor causes this vertex to become infected with probability 1. Unlike submodular contagions, nonsubmodular contagions can require well-connected regions to spread [9].

Influence maximization becomes qualitatively different in nonsubmodular settings. In the submodular case, one should put as much distance between the k initial adopters as possible, lest they erode each other’s effectiveness. However, in the nonsubmodular case, it may be advantageous to place the initial adopters close together to create synergy and yield more adoptions. The intuition that it is better to saturate one market first, and then expand implicitly assumes nonsubmodular influence in the cascades.

Key Question: Can this worst-case inapproximability result of $N^{1-\epsilon}$ for nonsubmodular influence maximization be circumvented by making realistic assumptions about either the underlying network topology or the cascade model?

We know a lot about what social networks look like, and previous hardness reductions make no attempt to capture realistic features of networks. It is very plausible that by restricting the space of networks we might regain tractability.

In this paper, we consider two natural network topologies: the hierarchical block model and the stochastic hierarchical blockmodel. Each is a natural restriction on the classic (*stochastic*) *blockmodel* [16, 22, 39] network structure. In (stochastic) blockmodels, agents are partitioned into ℓ blocks. The weight (or likelihood in the stochastic setting) of an edge between two vertices is based solely on blocks to which the vertices belong. The weights (or probabilities) of edges between two blocks can be represented by an $\ell \times \ell$ matrix. In the (stochastic) hierarchical blockmodel, the structure of the $\ell \times \ell$ matrix is severely restricted to be “tree-like”.¹

Our (stochastic) hierarchical blockmodel describes the hierarchical structure of the communities, in which a community is divided into many sub-communities, and each sub-community is further divided, etc. Typical examples include the structure of a country, which is divided into many provinces, and each province can be divided into cities. Our model captures the natural observation that people in the same sub-community in the lower hierarchy tend to have tighter (or more numerous) bonds among each other [13]. Such a highly abstracted model necessarily fails to capture all features of social networks. However, when we use this model as a lower bound, that is actually a strength as it shows that the problem is hard even in the case that communities structure can be represented by a tree. Additionally, we feel that this is a very natural model which captures salient features of real-world networks, so our upper bounds in this model are still interesting.

We also consider restrictions on the cascade model. The same research showing that cascades

¹Previous work on community detection in networks [29] defines a different, but related stochastic hierarchical blockmodel, where the hierarchy is restricted to two levels.

are often not submodular empirically also shows that the local submodularity often fails in one particular way—the second infected neighbor of an agent is, on average, more influential than the first. This has already been observed in community formation [3], viral marketing [27] and Twitter network [35]. This motivates our study of the *2-submodular* cascade model where the marginal effect of the second infected neighbor is greater than the first, but after that the marginal effect decreases.

1.1 Our Results

First, we present inapproximability results for INFMAX in both the hierarchical blockmodel and the stochastic hierarchical blockmodel. We show that INFMAX is NP-hard to approximate within a factor of $N^{1-\varepsilon}$ for arbitrary $\varepsilon > 0$. Moreover, this result holds in the hierarchical blockmodel even if we assume all agents have unit threshold $\theta_v = 1$. We also extend this hardness result to the stochastic hierarchical blockmodel.

Moreover, for the hierarchical blockmodel, we present a dynamic program based polynomial time algorithm for INFMAX when we additionally assume the influence from one block to another is “one-way”. This provides insights to the above intractability result: the difficulty comes from the bidirectionality of influence between agent-blocks.

Secondly, we present an inapproximability result for the 2-submodular cascade model. In particular, for *any* 2-submodular influence function f , we show that it is NP-hard to approximate INFMAX within a factor of N^τ when each agent has f as its local influence function, where $\tau > 0$ is a constant depending on f . This can be seen as a threshold result for approximability of INFMAX, because if f is submodular, then the problem can be approximated to within a $(1 - 1/e)$ -factor, but if f is just barely nonsubmodular the problem can no longer be approximated to within any constant factor.

Finally, we pose the open question of whether enforcing the aforementioned restrictions simultaneously on the network and the cascade renders the problem tractable.

1.2 Related Work

The influence maximization problem was posed by Domingos and Richardson [17, 34]. Kempe, Kleinberg, and Tardos showed that a simple greedy algorithm obtains a $(1 - 1/e)$ factor approximation to the problem in the independent cascade model and linear threshold model [24], and extended this result to a family of submodular cascades which captures the prior results as a special case [25]. Mossel and Roch [32] further extended this result to capture all submodular cascades.

Perhaps most related to the present work, are several inapproximability results for INFMAX. If no assumption is made for the influence function, INFMAX is NP-hard to approximate to within a factor of $N^{1-\varepsilon}$ for any $\varepsilon > 0$ [25].

Chen [10] found inapproximability results on a similar optimization problem: instead of maximizing the total number of infected vertices given k initial targets, he considered the problem of finding a minimum-sized set of initial seeds such that all vertices will eventually be infected. This work studied restrictions of this problem to various threshold models.

An important difference between our hardness result in Section 6 and all the previous results is that our result holds for *any* 2-submodular functions. In particular, in this work, f is fixed in advance before the NP-hardness reduction, while in previous work, specific influence functions were constructed within the reductions.

Several works looked at slightly different aspects of influence maximization. Borgs, Brautbar, Chayes, and Lucier [7] provably showed fast running times when the influence function is the independent cascade model. Lucier, Oren, and Singer [28] showed how to parallelize (in a model based on Map Reduce) the subproblem of determining the influence of a particular seed. Seeman and Singer [36] studied the special case where only a subset of the nodes in the network are available to be infected. They showed a constant factor approximation to the problem in their setting. He and Kempe looked at a robust versions of the problem [21] where the exact parameters of the cascade are unknown. Several works [6, 19] studied the problem as a game between two different infectors.

Following the work of Kempe, Kleinberg, and Tardos [24, 25], there were extensive works to solve INFMAX based on the heuristic implementations of the greedy algorithm designed to be efficient and scalable [11, 12, 28].

The notion of “near submodularity” was also proposed and studied in [37]. Our definition differs from the one in [37] in that a 2-submodular function can be, intuitively, very far from being submodular (for example, the 2-threshold cascade model). However, our reduction in Section 6 works for all 2-submodular functions, and 2-submodular functions can be arbitrarily close to submodular functions.

Our algorithm in Section 5 was further studied and generalized by Angell and Schoenebeck in [1]. They showed that, empirically, this generalized algorithm works very well even for arbitrary graphs. Specifically, they run dynamic programming on a hierarchical decomposition of general graphs, and, empirically, the algorithm effectively leverages the resultant hierarchical structures to return seed sets substantially superior to those of the greedy algorithm.

2 Preliminaries

In general a *cascade* on a graph is a stochastic mapping from a subset of vertices—the *seed vertices*, to another set of vertices that always contain the seed vertices—the *infected vertices*. The cascades we study in this paper all belong to the general threshold model [32], which captures the local decision making of vertices.

Definition 1. *The **general threshold model** $I_{F,\mathcal{D}}^G$, is defined by a graph $G = (V, E)$ which may or may not be edge-weighted, and for each vertex v :*

- i. a monotone local influence function $f_v : \{0, 1\}^{|\Gamma(v)|} \mapsto \mathbb{R}_{\geq 0}$ where $\Gamma(v)$ denotes the neighbor vertices of v and $f_v(\emptyset) = 0$, and*
- ii. a threshold distribution \mathcal{D}_v whose support is $\mathbb{R}_{\geq 0}$. Let F and \mathcal{D} denote the collection of f_v and \mathcal{D}_v respectively.*

On input $S \subseteq V$, $I_{F,\mathcal{D}}^G(S)$ outputs a set of vertices as follows:

- 1. Initially only vertices in S are infected, and for each vertex v the threshold $\theta_v \sim \mathcal{D}_v$ is sampled from \mathcal{D}_v independently.²*
- 2. In each subsequent round, a vertex v becomes infected if the influence of its infected neighbors exceeds its threshold.*
- 3. The set of infected vertices is the output (after a round where no additional vertices are infected).*

²The rationale of sampling thresholds *after* the seeds selection is to capture the scenario that the seed-picker does not have the full information on the agents in a social network, and this setting has been used in many other works [24, 32].

We use k to denote $|S|$ —the number of seeds, and use N to denote $|V|$ —the total number of vertices in G . Let

$$\sigma_{F,\mathcal{D}}^G(S) = \mathbb{E} [|I_{F,\mathcal{D}}^G(S)|]$$

be the *expected* total number of infected vertices due to the influence of S , where the expectation is taken over the samplings of the thresholds of all vertices. We refer to $\sigma_{F,\mathcal{D}}^G(\cdot)$ as the *global influence function*. Sometimes we write $\sigma(\cdot)$ with the parameters G, F, \mathcal{D} omitted, when there is no confusion. Because each f_v is monotone, it is straightforward to see that σ is monotone.

Definition 2. *The INFMAX problem is an optimization problem which takes as inputs $G = (V, E)$, F, \mathcal{D} , and an integer k , and outputs $\max_{S \subseteq V: |S|=k} \sigma_{F,\mathcal{D}}^G(S)$, the maximum global influence of a set of size k .*

In this paper, we consider several special cases of the general threshold model $I_{F,\mathcal{D}}^G$ by making assumptions on the network topology G , or the cascade model³ F, \mathcal{D} .

2.1 Assumptions on Graph G

We consider two graph models—the *hierarchical blockmodel* and the *stochastic hierarchical blockmodel*, which are the special case of the well studied *blockmodel* [39] and *stochastic blockmodel* [22] respectively.

The Hierarchical Blockmodel

Definition 3. *A **hierarchical blockmodel** is a undirected **edge-weighted** graph $G = (V, T)$, where V is the set of all vertices of the graph G , and $T = (V_T, E_T, w_T)$ is a **node-weighted binary tree** T called a **hierarchy tree**. In addition, w_T satisfies $w_T(t_1) \leq w_T(t_2)$ for any $t_1, t_2 \in V_T$ such that t_1 is an ancestor of t_2 .⁴ Each leaf node $t \in V_T$ corresponds to a subset of vertices $V(t) \subseteq V$, and the $V(t)$ sets partition the vertices of V . In general, if t is not a leaf, we denote $V(t) = \cup_{t'}: \text{a leaf, and an offspring of } t V(t')$.*

For $u, v \in V$, the weight of the edge (u, v) in G is just the weight of the least common ancestor of u and v in T . That is $w(u, v) = \max_{t: u, v \in V(t)} w(t)$. If this weight is 0, then we say that the edge does not exist.

To avoid possible confusion, we use the words *node* and *vertex* to refer to the vertices in T and G respectively.

Figure 1 provides an example of how a hierarchy tree defines the weights of edges in the corresponding graph.

Additionally, we can assume without loss of generality that the hierarchy tree is a *full* binary tree, as a node in T having only one child plays no role at deciding the weights of edges in G . For example, in Figure 1, the node having weight 2 does not affect the weight configuration on the right hand side. We can delete this node and promote the node with weight 5 to be a child of the root node. We will keep the full binary tree assumption from now on.

³The phrase “cascade model” here, as well as in the abstract and Section 1, refers to the description how each vertex is influenced by its neighbors, which is completely characterized by F and \mathcal{D} in the general threshold model.

⁴Since, as it will be seen later, each node in the hierarchy tree represents a community and its children represent its sub-communities, naturally, the relation between two persons is stronger if they are in a same sub-community in a lower level.

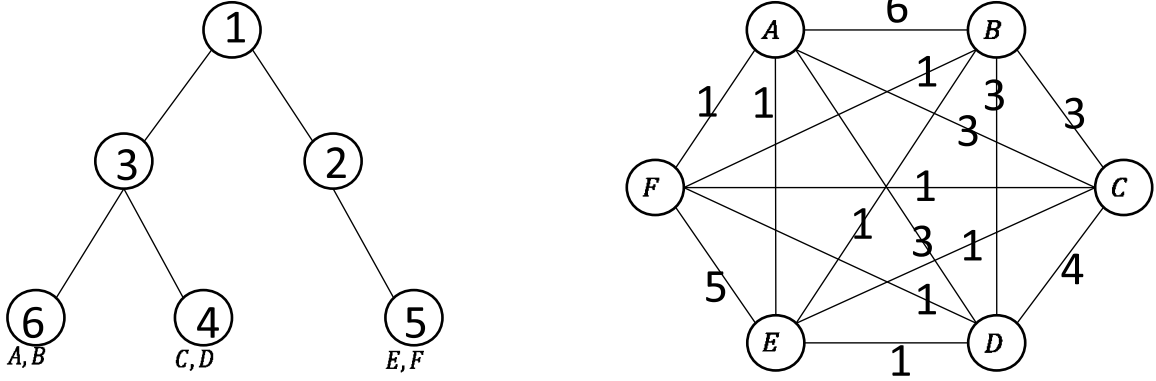


Figure 1: An example of a hierarchy tree with its corresponding graph. The number on each node of the hierarchy tree on the left hand side indicates the weight of the node, which reflects the weight of the corresponding edges on the hierarchical block graph on the right hand side in the above-mentioned way.

The Stochastic Hierarchical Blockmodel The *stochastic hierarchical blockmodel* is similar to the hierarchical blockmodel defined in the last section, in the sense that the structure of the graph is determined by a hierarchy tree. Instead of assigning weights to different edges measuring the strength of relationships, here we assign a probability with which the edge between each pair of vertices appears. Technically speaking, a stochastic hierarchical blockmodel is a distribution of unweighted undirected graphs, where each edge is sampled with a certain probability.

Definition 4. A *stochastic hierarchical blockmodel* is a distribution $\mathcal{G} = (V, T)$ of unweighted undirected graphs where V, T are the same as they are in Definition 3 with the additional restriction that the node weights in T belong to the interval $[0, 1]$. Let H be the weighted graph defined by the hierarchical blockmodel $H = (V, T)$, and let $w(e)$ denote the weight of edge e in H . Then $G = (V, E)$ is sampled by independently including each edge e with probability $w(e)$.

When it comes to the choices of S , the INFMAX problem can be defined in two different ways, regarding whether or not we allow the seed-picker to see the sampling $G \sim \mathcal{G}$ before choosing the seed set S .

Definition 5. *Pre-sampling stochastic hierarchical blockmodel* INFMAX is an optimization problem which takes as inputs $\mathcal{G}, F, \mathcal{D}$, an integer k and outputs $\max_{S \subseteq V: |S|=k} \mathbb{E}_{G \sim \mathcal{G}} [\sigma_{F, \mathcal{D}}^G(S)]$, the maximum expected global influence of a set of size k .

Definition 6. *Post-sampling stochastic hierarchical blockmodel* INFMAX is an average case version of INFMAX which takes as input $\mathcal{G}, F, \mathcal{D}$, and an integer k , and outputs the INFMAX Instance (G, F, \mathcal{D}, k) after sampling G from \mathcal{G} .

2.2 Assumptions on Cascade Model F, \mathcal{D}

We consider several generalizations of the well-studied *linear threshold model* [24]. The linear threshold model is a special case of the general threshold model $I_{F, \mathcal{D}}^G$, with each f_v being linear (see Definition 7 below), and each \mathcal{D}_v being the uniform distribution on $[0, 1]$.

The cascade model in Definition 7 generalizes the linear threshold model by removing the assumption on \mathcal{D}_v . The *universal local influence model* defined in Definition 9, generalizes the linear

threshold model by allowing non-linear f_v , while it restricts our attention to unweighted graphs. We also consider a special case where f_v is *2-submodular* in the last subsection.

Linear and Counting Local Influence Functions A natural selection of local influence function f_v is the linear function, by which the influences from v 's neighbors are additive.

Definition 7. Given a general threshold model $I_{F,\mathcal{D}}^G$ with a weighted graph G , we say that F is *linear* if for each $v \in V$ we have $f_v(S_v) = \sum_{u \in S_v} w(u, v)$.

For a general threshold model $I_{F,\mathcal{D}}^G$ with linear F , if we additionally assume each \mathcal{D}_v is the uniform distribution on $[0, 1]$, then this becomes the linear threshold model.

Definition 7 defines a cascade model for weighted graphs. We have the following definition which is the unweighted counterpart to Definition 7.

Definition 8. Given a general threshold model $I_{F,\mathcal{D}}^G$ with a unweighted graph G , we say that F is *counting* if for each $v \in V$ we have $f_v(S_v) = |S_v|$.

Universal Local Influence Functions We say f_v is *symmetric* if $f_v(S_v)$ only depends on the number of v 's infected neighbors $|S_v|$ so that each of v 's infected neighbors is of equal importance. In this case, f_v can be viewed as a function $f_v : \mathbb{Z}_{\geq 0} \mapsto \mathbb{R}_{\geq 0}$ which takes an integer as input, rather than a set of vertices. Thus f_v can be encoded by an increasing sequence of positive real numbers a_0, a_1, a_2, \dots so that $f_v(i) = a_i$. Note that $f_v(0) = a_0 = 0$, as we have assumed $f_v(\emptyset) = 0$.

For instance, the local influence function f_v defined in Definition 8 is symmetric, with $a_i = i$. In contrast, f_v in Definition 7 is not symmetric, as the neighbors connected by heavier edges contribute more to $f_v(S_v)$.

Definition 9. Given an increasing function $f : \mathbb{Z}_{\geq 0} \mapsto [0, 1]$, the *universal local influence model* I_f^G is a special case of the general threshold model $I_{F,\mathcal{D}}^G$, such that for each $v \in V$ we have

- f_v is symmetric, and $f_v = f$ (such that all f_v 's are identical).
- \mathcal{D}_v is the uniform distribution on $[0, 1]$.

Notice that we can assume without loss of generality that G is unweighted in Definition 9, as each f_v is fixed to be some increasing function f which does not depend on the weights of edges.

After assuming G is unweighted, the universal local influence model is a generalization of the linear threshold model: the linear threshold model can be viewed as the universal local influence model by restricting $a_i = i$.

As a final remark, for any general threshold model $I_{F,\mathcal{D}}^G$ with each \mathcal{D}_v being the uniform distribution on $[0, 1]$, we can intuitively view $f_v(S_v)$ as the *probability* that v will be infected (where we take $f_v(S_v) > 1$ as probability 1). In the universal local influence model, a_i can be viewed as the probability that a vertex will be infected, given that it has i infected neighbors.

Submodular and 2-Submodular Functions Let $g : \{0, 1\}^S \mapsto \mathbb{R}$ be a function which takes as input a subset of a set S . Formally, g is *submodular* if $g(A \cup \{u\}) - g(A) \geq g(B \cup \{u\}) - g(B)$ for any $u \in S$ and sets $A \subseteq B \subseteq S$. Intuitively, this means that the marginal effect of each element decreases as the set increases.

The definition above can be applied to each local influence functions $f_v : \{0, 1\}^{|\Gamma(v)|} \mapsto \mathbb{R}_{\geq 0}$, as well as the global influence function $\sigma_{F,\mathcal{D}}^G : \{0, 1\}^{|V|} \mapsto \mathbb{R}_{\geq 0}$. Given G, F, \mathcal{D} we say that a

general threshold model $I_{F, \mathcal{D}}^G(\cdot)$ is submodular if $\sigma_{F, \mathcal{D}}^G(\cdot)$ is. In [32], it has been shown that the local submodularity of all f_v 's implies the global submodularity of $I_{F, \mathcal{D}}^G(\cdot)$ for all G when \mathcal{D}_v is the uniform distribution on $[0, 1]$.

We are particularly concerned with the universal local influence model in Definition 9. Here f is submodular if the marginal gain of f by having one more infected neighbor is non-increasing as the number of infected neighbors increases. Formally, for $i_1 < i_2$, we have

$$f(i_1 + 1) - f(i_1) \geq f(i_2 + 1) - f(i_2).$$

Intuitively, f is submodular if its domain can be smoothly extended to $\mathbb{R}_{\geq 0}$ to make f concave.

We will consider *2-submodular* local influence functions f , which is “almost” submodular such that the submodularity is only violated for the first two inputs of f . In particular, we fail to have the submodular constraint $f(1) - f(0) \geq f(2) - f(1)$, and instead we have $f(1) - f(0) < f(2) - f(1)$, which is just $f(2) > 2f(1)$ as $f(0) = 0$.

Definition 10. $f : \mathbb{Z}_{\geq 0} \mapsto [0, 1]$ is **2-submodular** if $f(2) > 2f(1)$ and $f(i) - f(i - 1)$ is non-increasing in i for $i \geq 2$.

In general, for any non-zero submodular function f , if we sufficiently decrease $f(1)$, f becomes 2-submodular. Thus, from any non-zero submodular function, we can obtain a 2-submodular function.

We note that the 2-threshold cascade model, where each vertex will be infected if it has at least 2 infected neighbors, can be viewed as the universal local influence model with a 2-submodular f (with $f(0) = f(1) = 0$ and $f(i) = 1$ for $i \geq 2$, keeping the assumption that θ_v is drawn uniformly at random from $[0, 1]$).

3 Hierarchical Blockmodel Influence Maximization

In this section, we provide a strong inapproximability result for INFMAX problem for the hierarchical blockmodel cascade even when all vertices have a deterministic threshold 1. Specifically, we will show that it is NP-hard to approximate optimal $\sigma(S)$ within a factor of $N^{1-\varepsilon}$ for any $\varepsilon > 0$. The same inapproximability result holds for the most general case where \mathcal{D} is given as input to INFMAX.

Theorem 1. For any constant $\varepsilon > 0$, INFMAX (G, F, \mathcal{D}, k) is NP-hard to approximate to a factor of $N^{1-\varepsilon}$, even if G is a hierarchical blockmodel, F is linear (see Definition 7), and \mathcal{D}_v is the degenerate distribution with mass 1 on $\theta_v = 1$ for all $v \in V$.

We will prove Theorem 1 by a reduction from the VERTEXCOVER problem, a well-known NP-complete problem.

Definition 11. Given an undirected graph $\bar{G} = (\bar{V}, \bar{E})$ and a positive integer \bar{k} , the VERTEXCOVER problem (\bar{G}, \bar{k}) asks if we can choose a subset of vertices $\bar{S} \subseteq \bar{V}$ such that $|\bar{S}| = \bar{k}$ and such that each edge is incident to at least one vertex in \bar{S} .

The Reduction Given a VERTEXCOVER instance (\bar{G}, \bar{k}) , let $n = |\bar{V}|$ and $m = |\bar{E}|$. We use A_1, \dots, A_n to denote the n vertices and e_1, \dots, e_m to denote the m edges.⁵ We make the assumptions

⁵We use the letter A to denote the vertices in a VERTEXCOVER instance instead of commonly used v , while v is used for the vertices in an INFMAX instance. Since VERTEXCOVER can be viewed as a special case of SETCOVER with vertices corresponding to subsets and edges corresponding to elements, the letter A , commonly used for subsets, is used here.

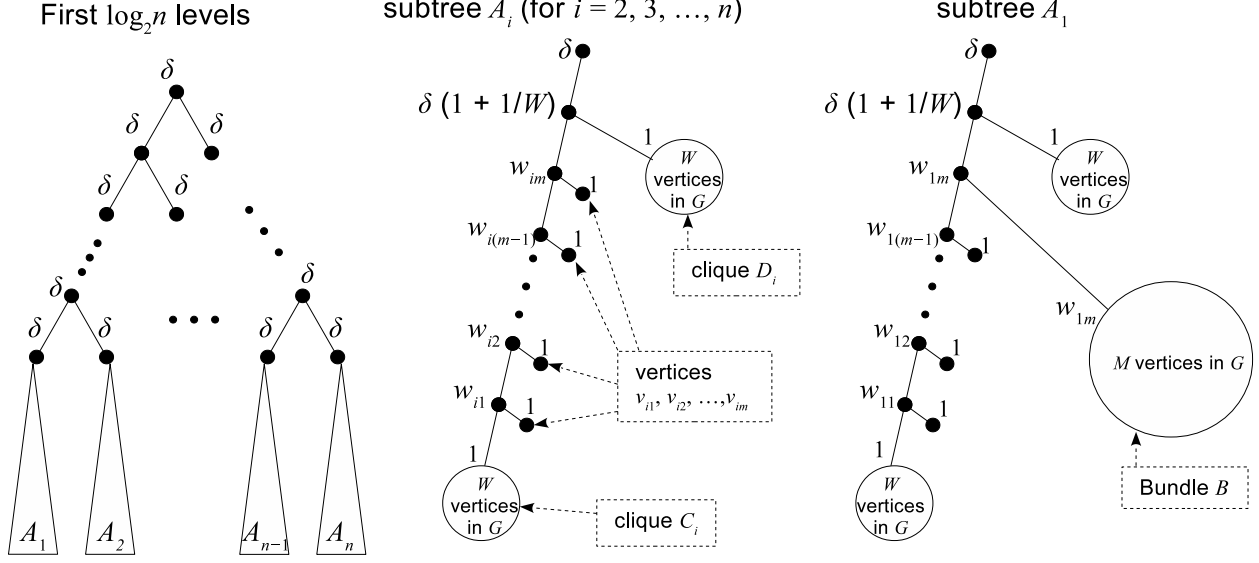


Figure 2: The construction of the hierarchy tree T .

$n > \bar{k}$ and $m > n + \bar{k}$.⁶ Let $W = nm$, $M = (n(2W + m) - 1)^{\frac{1}{\varepsilon}}$, and $\delta > 0$ be a sufficiently small real number.

We will construct the graph $G = (V, E, w)$ by constructing a hierarchy tree T which uniquely determines G (see Definition 3 in Section 2.1). The construction of T is shown in Figure 2. The first $\log_2 n$ levels of T is a full balanced binary subtree with n leaves, and the weight of the nodes in all these levels is δ . Each of those n leaves is the root of a subtree corresponding to each vertex A_i in the VERTEXCOVER instance.

The structure of the subtrees corresponding to A_2, \dots, A_n and A_1 are shown on the right hand side of Figure 2. The numbers on the tree nodes indicate the weights, and in particular

$$w_{ij} = \begin{cases} \frac{[1 - (n + \bar{k} - 1)W\delta - (n - 1)(j - 1)\delta - 2\delta] + \delta}{W - 1 + j} & \text{if edge } e_j \text{ is incident to } A_i \\ \frac{1 - (n + \bar{k} - 1)W\delta - (n - 1)(j - 1)\delta - 2\delta}{W - 1 + j} & \text{otherwise} \end{cases},$$

for each $i = 1, \dots, n$ and $j = 1, \dots, m$.

The leaves of each subtree A_i are the leaves of T , which, as we recall from Definition 3 correspond to subsets of vertices in $G = (V, E, w)$. Among all the leaves shown on the right hand side of Figure 2, each solid dot corresponds to a subset of V containing only one vertex, and each hollow circle corresponds to a subset of V containing many vertices with the corresponding number of vertices shown.

For each subtree A_i with $i = 2, \dots, n$, we have constructed $m + 2$ leaves corresponding to $2W + m$ vertices in G . They are, in up-to-down order, a clique D_i of W vertices, vertices $v_{im}, v_{i(m-1)}, \dots, v_{i1}$, and a clique C_i of W vertices. As each vertex has threshold 1 and the leaf nodes corresponding to C_i, D_i both have weight 1, infecting any vertex in C_i or D_i will cause the infection of all W vertices (which justifies the name ‘‘clique’’).

The construction of A_1 is similar. The only difference is that, instead of connecting to a node corresponding to the vertex v_{1m} , the node with weight w_{1m} is now connected to another node with

⁶For the assumption $m > n + \bar{k}$, notice that allowing the graph \bar{G} to be a multi-graph does not change the nature of VERTEXCOVER, we can ensure m to be sufficiently large by just duplicating edges.

the same weight and corresponding to a bundle B in G with M vertices. We shall not call this large bundle B a “clique”, as the weight of the edge between each pair of these M vertices is $w_{1m} \ll 1$, which is much weaker.

It is easy to calculate the total number of vertices in the construction: $N = M + M^\varepsilon$.

The Reduction Correctness For a VERTEXCOVER instance (\bar{G}, \bar{k}) , consider the INFMAX instance (G, F, \mathcal{D}, k) with $k = n + \bar{k}$. We aim to show that,

1. If the VERTEXCOVER instance (\bar{G}, \bar{k}) is a YES instance, then there exists $S \subseteq V$ with $|S| = k$ such that $\sigma(S) \geq M$;
2. If the VERTEXCOVER instance (\bar{G}, \bar{k}) is a NO instance, then for any $S \subseteq V$ with $|S| = k$ we have $\sigma(S) \leq M^\varepsilon = n(2W + m) - 1$.

Proof of 1. Suppose we have a YES VERTEXCOVER instance (\bar{G}, \bar{k}) with $\bar{S} \subseteq \bar{V}$ covering all edges in \bar{E} . In the INFMAX instance, we aim to show that at least M vertices will be infected if we choose those $k = n + \bar{k}$ seeds in the following way:

- choose an arbitrary seed in each of the cliques C_1, \dots, C_n (a total of n seeds are chosen);
- for each $A_i \in \bar{S}$, choose an arbitrary seed in the clique D_i (a total of \bar{k} seeds are chosen).

By such a choice, in the first round of the cascade, all the W vertices in each of C_1, \dots, C_n and each of those \bar{k} (D_i)’s are infected. We aim to show that all vertices in B will be infected after at most $3m$ cascade rounds. We call the set of n vertices $\{v_{1j}, \dots, v_{nj}\}$ the j -th level, and we will show that the cascade carries on level by level. In particular, we will first show that all vertices in the first level will be infected in at most 3 rounds. Next, given that all vertices in the first j levels are infected, by similar calculations, we can show that all vertices in the $(j+1)$ -th level will be infected.

Consider the first level $\{v_{11}, \dots, v_{n1}\}$. Let $e_1 = (A_{i_1}, A_{i'_1}) \in \bar{E}$. Since the VERTEXCOVER instance is a YES instance, either $A_{i_1} \in \bar{S}$ or $A_{i'_1} \in \bar{S}$, or both. Assume $A_{i_1} \in \bar{S}$ without loss of generality, then all vertices in D_{i_1} are already infected. In the coming round, the vertex $v_{i_11} \in V$ will be infected, as

$$\begin{aligned} f_{v_{i_11}} \left(\bigcup_{i=1}^n C_i \cup \bigcup_{A_i \in \bar{S}} D_i \right) &= \delta \left| \bigcup_{i \neq i_1} C_i \cup \bigcup_{i \neq i_1, A_i \in \bar{S}} D_i \right| + w_{i_11} |C_{i_1}| + \delta \left(1 + \frac{1}{W} \right) |D_{i_1}| \\ &= \delta((n-1) + (\bar{k}-1))W + \frac{1 - (n + \bar{k} - 1)W\delta - \delta}{W} \cdot W + \delta \left(1 + \frac{1}{W} \right) W \\ &= 1. \end{aligned}$$

If $A_{i'_1} \in \bar{S}$ as well, then $v_{i'_11} \in V$ will also be infected in the this round, due to the same calculation. On the other hand, if $A_{i'_1} \notin \bar{S}$, $v_{i'_11}$ will be infected in the next round, as

$$\begin{aligned} f_{v_{i'_11}} \left(\bigcup_{i=1}^n C_i \cup \bigcup_{A_i \in \bar{S}} D_i \cup \{v_{i_11}\} \right) &= \delta \left| \bigcup_{i \neq i'_1} C_i \cup \bigcup_{A_i \in \bar{S}} D_i \cup \{v_{i_11}\} \right| + w_{i'_11} |C_{i'_1}| \\ &= \delta((n-1 + \bar{k})W + 1) + \frac{1 - (n + \bar{k} - 1)W\delta - \delta}{W} \cdot W \\ &= 1. \end{aligned}$$

Therefore, both v_{i_01} and $v_{i'_01}$ will be infected in both cases.

In the next round, the remaining $n - 2$ vertices $\{v_{i_01}\}_{i_0 \notin \{i_1, i'_1\}; 1 \leq i_0 \leq n}$ will be infected, as we have

$$\begin{aligned} f_{v_{i_01}} \left(\bigcup_{i=1}^n C_i \cup \bigcup_{A_i \in \bar{S}} D_i \cup \{v_{i_01}, v_{i'_01}\} \right) &= \delta \left| \bigcup_{i \neq i_0} C_i \cup \bigcup_{A_i \in \bar{S}} D_i \cup \{v_{i_01}, v_{i'_01}\} \right| + w_{i_01} |C_{i_0}| \\ &= \delta((n-1 + \bar{k})W + 2) + \frac{1 - (n + \bar{k} - 1)W\delta - 2\delta}{W} \cdot W \\ &= 1, \end{aligned}$$

in the case $A_{i_0} \notin \bar{S}$ (such that no vertex in D_{i_0} is infected at this moment), and

$$\begin{aligned} f_{v_{i_01}} \left(\bigcup_{i=1}^n C_i \cup \bigcup_{i \neq i_0, A_i \in \bar{S}} D_i \cup \{v_{i_01}, v_{i'_01}\} \right) &= \delta \left| \bigcup_{i \neq i_0} C_i \cup \bigcup_{i \neq i_0, A_i \in \bar{S}} D_i \cup \{v_{i_01}, v_{i'_01}\} \right| + w_{i_01} |C_{i_0}| + \delta \left(1 + \frac{1}{W}\right) |D_{i_0}| \\ &= \delta((n-1 + \bar{k} - 1)W + 2) + \frac{1 - (n + \bar{k} - 1)W\delta - 2\delta}{W} \cdot W + \delta \left(1 + \frac{1}{W}\right) W \\ &= 1 + \delta > 1, \end{aligned}$$

in the case $A_{i_0} \in \bar{S}$ (such that all vertices in D_{i_0} are infected at the first round). In conclusion, all the n vertices $\{v_{i_1}\}_{1 \leq i \leq n}$ will be eventually infected in at most 3 rounds.

The analysis of the second level is similar. For $e_2 = (A_{i_2}, A_{i'_2}) \in \bar{E}$, we have either $A_{i_2} \in \bar{S}$ or $A_{i'_2} \in \bar{S}$ (or both), making one of $v_{i_22}, v_{i'_22}$ infected (or both), which further makes both $v_{i_22}, v_{i'_22}$ infected (if one of them is not infected previously), and which eventually makes all the n vertices $\{v_{i_2}\}_{1 \leq i \leq n}$ infected.

For each $j = 1, \dots, m$ with $e_j = (A_{i_j}, A_{i'_j})$, we have either $A_{i_j} \in \bar{S}$ or $A_{i'_j} \in \bar{S}$ (or both). Similar as above, after either two or three rounds, all the vertices in $\{v_{i_j}\}_{1 \leq i \leq n}$ will be infected, if all the vertices in $\{v_{i_1}\}_{1 \leq i \leq n}, \dots, \{v_{i_{(j-1)}}\}_{1 \leq i \leq n}$ are already infected.

Therefore, we can see that the cascade after the first round carries on in the following order:

$$v_{i_11} \rightarrow v_{i'_11} \rightarrow \{v_{i_1}\}_{i \neq i_1, i'_1} \rightarrow v_{i_22} \rightarrow v_{i'_22} \rightarrow \{v_{i_2}\}_{i \neq i_2, i'_2} \rightarrow \dots \rightarrow v_{i_m m} \rightarrow v_{i'_m m} \rightarrow \{v_{i_m}\}_{i \neq i_m, i'_m} \rightarrow B.$$

Therefore, we conclude 1 as we already have M infected vertices by just counting those in the bundle B . \square

For the proof of 2, we present a general proof idea before the formal proof.

To show 2 by contradiction, we assume that we can choose a seed set $S \subseteq V$ such that $|S| = k = n + \bar{k}$ and $\sigma(S) > M^\varepsilon$. By a careful analysis, we can conclude that the only possible way to choose S is as follow.

- an arbitrary vertex from each of C_1, \dots, C_n (a total of n vertices are chosen);
- an arbitrary vertex from each of $D_{\pi_1}, \dots, D_{\pi_{\bar{k}}}$ for certain $\{\pi_1, \dots, \pi_{\bar{k}}\} \subseteq \{1, \dots, n\}$ (a total of \bar{k} vertices are chosen).

The intuitive reason for this is the following: firstly, choosing k seeds among the $2n$ cliques $C_1, \dots, C_n, D_1, \dots, D_n$ is considerably more beneficial, as a seed would cause the infection of W vertices; secondly, if we cannot choose both C_i and D_i , it is always better to choose C_i because the weights w_{i1}, \dots, w_{im} are considerably larger than $\delta(1 + 1/W)$, if δ is set sufficiently small.

Since the VERTEXCOVER instance is a NO instance, there exists an edge $e_j = (A_{i_j}, A_{i'_j})$ such that no vertex in D_{i_j} and $D_{i'_j}$ is chosen as seed. By following similar analysis as in the proof of 1, we can see that the cascade would stop at the level $\{v_{ij}\}_{i=1, \dots, n}$, which concludes 2.

Proof of 2. Assume that we can choose seed set $S \subseteq V$ such that $|S| = k = n + \bar{k}$ and $\sigma(S) > M^\epsilon$. First notice that choosing any seeds from B is at most as good as choosing seeds from $C_1 \cup \{v_{1j}\}_{1 \leq j \leq m-1}$. By our assumption $m > n + \bar{k} = k$, we can assume without loss of generality that no seed is chosen in B . With this assumption, we will prove that none of these M vertices will be infected in the cascade. Since the graph G has a total of $N = M + M^\epsilon$ vertices, this contradicts to $\sigma(S) > M^\epsilon$.

Suppose, for the sake of contradiction, a vertex $u \in B$ is infected in round t of the cascade, and there is no infected vertex in B in the first $t - 1$ rounds. Let I_u be the set of infected vertices before round t . Since u is infected in round t , we have $f_u(I_u) \geq 1$, which, by Definition 3, implies

$$\sum_{v \in I_u} w(u, v) \geq 1.$$

We analyze the constituents of I_u .

We set δ to be sufficiently (but still polynomially) small such that

$$(n - 1)(2W + m)\delta + \delta \left(1 + \frac{1}{W}\right) \ll w_{1m}.$$

Then the infection of each vertex in $C_1 \cup \{v_{1j}\}_{1 \leq j \leq m-1}$ has contribution w_{1m} to $f_u(I_u)$, while the net contribution from the infections of all vertices in $V \setminus \{C_1 \cup \{v_{1j}\}_{1 \leq j \leq m-1} \cup B\}$ is much less than w_{1m} . On the other hand, even if all the $W + m - 1$ vertices in $C_1 \cup \{v_{1j}\}_{1 \leq j \leq m-1}$ are included in I_u , the contribution to $f_u(I_u)$ is

$$(W + m - 1)w_{1m} \leq 1 - (n + \bar{k} - 1)W\delta - (n - 1)(m - 1)\delta - \delta < 1,$$

which is still not enough. Thus, we conclude that $C_1 \cup \{v_{1j}\}_{1 \leq j \leq m-1} \subseteq I_u$, and the vertices from $V \setminus \{C_1 \cup \{v_{1j}\}_{1 \leq j \leq m-1} \cup B\}$ should contribute at least $(n + \bar{k} - 1)W\delta + (n - 1)(m - 1)\delta + \delta$ to $f_u(I_u)$. From the term $(n + \bar{k} - 1)W\delta$, we can see that at least $n + \bar{k} - 1$ cliques from the $2n - 1$ cliques $C_2, \dots, C_n, D_1, \dots, D_n$ must be included in I_u . Coupled with the observation $C_1 \subseteq I_u$, we need at least $n + \bar{k}$ infected cliques from $C_1, \dots, C_n, D_1, \dots, D_n$.

On the other hand, the only way to infect a clique C_i or D_i is to seed one of its vertices. To see this for each D_i , it is enough to notice that the weight $\delta(1 + 1/W)$ is extremely small. To see this for each C_i , notice that only v_{i1}, \dots, v_{im} have non-negligible influence to C_i , and

$$\sum_{j=1}^m w_{ij} < \sum_{j=1}^m \frac{1}{W - 1 + j} < m \times \frac{1}{W} = \frac{1}{n} \ll 1.$$

⁷This is always possible: when $\delta \rightarrow 0$, the left hand side approaches to 0, while we have $\lim_{\delta \rightarrow 0} w_{1m} = \frac{1}{W + m - 1}$ for the right hand side.

Therefore, in order to have $u \in B$ infected in round t , the only possible way is to choose $k = n + \bar{k}$ seeds from $n + \bar{k}$ cliques, among all the $2n$ cliques $C_1, \dots, C_n, D_1, \dots, D_n$. Lastly, it is straightforward to check that infecting any vertex in D_i is not as good as infecting a vertex in the corresponding C_i . Thus, we can assume without loss of generality that S consists of

- an arbitrary vertex from each of C_1, \dots, C_n (a total of n vertices are chosen);
- an arbitrary vertex from each of $D_{\pi_1}, \dots, D_{\pi_{\bar{k}}}$ for certain $\{\pi_1, \dots, \pi_{\bar{k}}\} \subseteq \{1, \dots, n\}$ (a total of \bar{k} vertices are chosen).

Since the VERTEXCOVER instance is a NO instance, for the choice $\bar{S} = \{A_{\pi_1}, \dots, A_{\pi_{\bar{k}}}\}$, there exists edge e_j that is not covered by \bar{S} . Let j^* be the smallest j such that e_j is not covered by \bar{S} .

We consider two cases: $j^* = 1$ and $j^* > 1$.

If $j^* = 1$, for $e_1 = (A_{i_1}, A_{i'_1})$, we have $A_{i_1}, A_{i'_1} \notin \bar{S}$. In this case, $v_{i_1 1}$ will not be infected, as

$$\begin{aligned} f_{v_{i_1 1}} \left(\bigcup_{i=1}^n C_i \cup \bigcup_{A_i \in \bar{S}} D_i \right) &= \delta \left| \bigcup_{i \neq i_1} C_i \cup \bigcup_{A_i \in \bar{S}} D_i \right| + w_{i_1 1} |C_{i_1}| \\ &= \delta(n-1 + \bar{k})W + \frac{1 - (n + \bar{k} - 1)W\delta - \delta}{W} \cdot W \\ &= 1 - \delta < 1, \end{aligned}$$

and $v_{i'_1 1}$ will not be infected for the same reason. For $i_0 \neq i_1, i'_1$, $v_{i_0 1}$ will not be infected either, as we have

$$\begin{aligned} f_{v_{i_0 1}} \left(\bigcup_{i=1}^n C_i \cup \bigcup_{A_i \in \bar{S}} D_i \right) &= \delta \left| \bigcup_{i \neq i_0} C_i \cup \bigcup_{A_i \in \bar{S}} D_i \right| + w_{i_0 1} |C_{i_0}| \\ &= \delta(n-1 + \bar{k})W + \frac{1 - (n + \bar{k} - 1)W\delta - 2\delta}{W} \cdot W \\ &= 1 - 2\delta < 1, \end{aligned}$$

in the case $A_{i_0} \notin \bar{S}$, and

$$\begin{aligned} f_{v_{i_0 1}} \left(\bigcup_{i=1}^n C_i \cup \bigcup_{A_i \in \bar{S}} D_i \right) &= \delta \left| \bigcup_{i \neq i_0} C_i \cup \bigcup_{i \neq i_0, A_i \in \bar{S}} D_i \right| + w_{i_0 1} |C_{i_0}| + \delta \left(1 + \frac{1}{W} \right) |D_{i_0}| \\ &= \delta(n-1 + \bar{k} - 1)W + \frac{1 - (n + \bar{k} - 1)W\delta - 2\delta}{W} \cdot W + \delta \left(1 + \frac{1}{W} \right) W \\ &= 1 - \delta < 1, \end{aligned}$$

in the case $A_{i_0} \in \bar{S}$. Thus, none of $\{v_{i 1}\}_{1 \leq i \leq n}$ will be infected. Since $w_{ij_1} > w_{ij_2}$ whenever $j_1 < j_2$ for any i (easy to see by observing $w_{ij} \approx \frac{1}{W-1+j}$), none of $\{v_{ij}\}_{1 \leq i \leq n; 2 \leq j \leq m}$ will be infected. In particular, no vertex in B can be infected, which leads to the desired contradiction.

If $j^* > 1$, by the similar analysis in the proof of 1 for the YES instance case, after many cascade rounds, all vertices in $\{v_{ij}\}_{1 \leq i \leq n; 1 \leq j \leq j^* - 1}$ will be infected. For $e_{j^*} = (A_{i_{j^*}}, A_{i'_{j^*}})$, we have

$A_{i_{j^*}}, A_{i'_{j^*}} \notin \bar{S}$. In this case, $v_{i_{j^*}j^*}$ will not be infected, as

$$\begin{aligned}
& f_{v_{i_{j^*}j^*}} \left(\bigcup_{i=1}^n C_i \cup \bigcup_{A_i \in \bar{S}} D_i \cup \{v_{ij}\}_{1 \leq i \leq n; 1 \leq j \leq j^*-1} \right) \\
&= \delta \left| \bigcup_{i \neq i_{j^*}} C_i \cup \bigcup_{A_i \in \bar{S}} D_i \right| + w_{i_{j^*}j^*} \left| C_{i_{j^*}} \cup \{v_{i_{j^*}j}\}_{1 \leq j \leq j^*-1} \right| + \delta \left| \{v_{ij}\}_{i \neq i_{j^*}, 1 \leq j \leq j^*-1} \right| \\
&= \delta(n-1 + \bar{k})W + \frac{1 - (n + \bar{k} - 1)W\delta - (n-1)(j^*-1)\delta - \delta}{W-1+j^*} \cdot (W + j^* - 1) + \delta(n-1)(j^*-1) \\
&= 1 - \delta < 1,
\end{aligned}$$

and $v_{i'_{j^*}j^*}$ will not be infected for the same reason. Following similar analysis, $v_{i_0j^*}$ will not be infected for $i_0 \neq i_{j^*}, i'_{j^*}$, and none of $\{v_{ij^*}\}_{1 \leq i \leq n}$ will be infected. By the same observation $w_{ij_1} > w_{ij_2}$ whenever $j_1 < j_2$, none of $\{v_{ij}\}_{1 \leq i \leq n; j^* \leq j \leq m}$ will be infected. In particular, no vertex in B can be infected, which again leads to the desired contradiction. We conclude 2 here. \square

By 1 and 2, the INFMAX problem for G we have constructed is NP-hard to approximate within a factor of at least

$$\frac{M}{M^\varepsilon} = M^{1-\varepsilon} = \Theta(N^{1-\varepsilon}),$$

as $N = M + M^\varepsilon = \Theta(M)$. Since ε is arbitrary, the inapproximability factor can be written as just $N^{1-\varepsilon}$.

4 Stochastic Hierarchical Blockmodel Influence Maximization

In this section, we will present strong inapproximability results for both pre-sampling and post-sampling versions of stochastic hierarchical blockmodel INFMAX. A major difference between the results in Section 3 and this section is that the strong inapproximability result no longer holds if we assume $\theta_v = 1$ for all $v \in V$ in the stochastic hierarchical blockmodel. In fact, if all the thresholds are fixed to be 1 and F is counting (see Definition 8), $\sigma(\cdot)$ in both Definition 5 and Definition 6 become submodular, in which case we can have a simple greedy $(1 - 1/e)$ -approximation algorithm [24, 33]. In particular, assuming $\theta_v = 1$ for all $v \in V$ makes post-sampling INFMAX trivial: as an infected seed will eventually infect a whole connected component of G , the optimal way of choosing S is to choose k seeds from the first k largest connected components, after seeing the sampling $G \sim \mathcal{G}$. For pre-sampling INFMAX, the model becomes the *independent cascade model* [24], which is known to be submodular.

The following two theorems are the same, except that Theorem 2 corresponds to the hardness for pre-sampling model (see Definition 5), while Theorem 3 show the same hardness result for the post-sampling model (see Definition 6) via a randomized Karp's reduction.

Theorem 2. *For any $\varepsilon > 0$, pre-sampling stochastic hierarchical blockmodel INFMAX is NP-hard to approximate within a factor of $N^{1-\varepsilon}$, even if F is counting and \mathcal{D}_v is a degenerated distribution on certain integer θ_v for each $v \in V$.*

Theorem 3. *For any $\varepsilon > 0$ and $c > 0$, approximating post-sampling stochastic hierarchical blockmodel INFMAX to within a factor of $N^{1-\varepsilon}$ with probability at least N^{-c} is NP-hard, even if F is counting and \mathcal{D}_v is a degenerated distribution on certain integer θ_v for each $v \in V$.*

As a remark to Theorem 3, the theorem says that if we have an oracle that outputs a solution which approximates $\max_{S \subseteq V, |S| \leq k} \sigma(S)$ within a factor of $N^{1-\varepsilon}$ for certain samples $G \sim \mathcal{G}$, and with probability at least N^{-c} we receive a sample G in the set of graphs for which the oracle outputs valid solutions, then we can use this oracle to solve any NP-complete problem as long as we have randomness to sample $G \sim \mathcal{G}$.

We will prove both Theorem 2 and Theorem 3 by a reduction from VERTEXCOVER. Given a VERTEXCOVER instance $(\bar{G} = (\bar{V}, \bar{E}), \bar{k})$, we will construct a hierarchy tree T which determines \mathcal{G} for both proofs.

The Reduction Let $n = |\bar{V}|$ and $m = \bar{E}$ as usual. Assume $m > n > \bar{k}^2 + 2$, and $\log_2 n$ is an integer.⁸ In addition, we assume that $A_1 \in \bar{S}$ whenever the VERTEXCOVER instance is a YES instance.⁹

We define the following variables used in this section.

$$\delta = \frac{1}{10mn^2\bar{k}}, \quad \text{and} \quad \Delta = mn^2\delta = \frac{1}{10\bar{k}}, \quad W = m^{10}n^{10}.$$

Let M be an extremely large number whose value will be decided later.

The construction of T is shown in Figure 3. T is a full balanced binary tree with $\log_2 n$ levels and n leaves. The weight of all non-leaf nodes is $1/W$, and the weight of all leaves is 1. The i -th leaf corresponds to $A_i \in \bar{V}$ in the VERTEXCOVER instance. Recall from Definition 4 that $\mathcal{G} = (V, T)$ is determined by T , and in particular each leaf of T corresponds to a subset of V . As the weight of each leaf is 1, meaning each edge appear with probability 1, its corresponding subset of vertices form a clique in all $G \sim \mathcal{G}$. We will call the clique corresponding to the i -th leaf *the i -th clique* in the remaining part of this section. For each clique i , we will first describe the vertices we have constructed in Figure 3, and then define their thresholds.

For positive integers x, y , denote by $B(x, y)$ a bundle of x vertices with threshold y . For each $i = 1, \dots, n$, we construct the following vertices for the i -th clique:

- a bundle of $\bar{k}W^2$ vertices: $B_i := B(\bar{k}W^2, \infty)$, and
- $m(n-2)$ bundles of W^3 vertices: $B_{ij\iota} := B(W^3, \theta_{ij\iota})$ for $j = 1, \dots, m$ and $\iota = 1, \dots, n-2$.

For $i = 1$, we add an extra bundle $C := (M, \theta_{1(m+1)})$. The thresholds $\{\theta_{ij\iota}\}$ and $\theta_{1(m+1)}$ of those constructed vertices will be defined later.

By our construction, the 1-st clique has $M + \bar{k}W^2 + m(n-2)W^3$ vertices, which is much more than the number of vertices $\bar{k}W^2 + m(n-2)W^3$ in each of the remaining cliques. As a remark, we have constructed $N = M + nm(n-2)W^3 + n\bar{k}W^2$ vertices for G . Moreover, for M whose value we have not decided yet, we can make it arbitrarily close to N .

Denote by $B_{\cdot j\iota} := \{B_{ij\iota}\}_{i=1, \dots, n}$ the n bundles in a horizontal level in Figure 3 (for example, in Figure 3, after the top level $\{B_1, \dots, B_n\}$, there come levels $B_{\cdot 11}, B_{\cdot 12}, \dots$). We will call $B_{\cdot j\iota}$ a *level* and abuse the word “level” to refer to the vertices in $B_{\cdot j\iota}$.

The correspondence between the VERTEXCOVER instance and the graph we constructed is as follows. Recall that each vertex $A_i \in \bar{V}$ corresponds to the i -th clique. Now, for each edge $e_j \in \bar{E}$, we have constructed $n-2$ levels $B_{\cdot j1}, \dots, B_{\cdot j(n-2)}$, which are $n(n-2)$ bundles of W^3 vertices. For

⁸Notice that we can assume $n \gg \bar{k}$ is an integer power of 2 by adding isolated vertices to \bar{G} which are never picked, and we can assume $m > n$ by duplicate each edge (which makes \bar{G} a multi-graph).

⁹This assumption can be made without loss of generality because we can add two extra vertices named A_1, A_2 and one extra edge (A_1, A_2) such that one of A_1, A_2 must be chosen to cover this edge, and we can assume A_1 is chosen.

example, in Figure 3, we have illustrated the $n - 2$ levels corresponding to e_1 and the $n - 2$ levels corresponding to e_m , while the levels corresponding to the remaining edges in \bar{E} are omitted.

For each $j = 1, \dots, m$ and $\iota = 1, \dots, n - 2$, we denote by $B_{\prec j \iota}$ the union of the first $(j - 1)(n - 2) + \iota - 1$ levels (where the levels are ordered from up to down in Figure 3):

$$\begin{aligned} B_{\prec j \iota} &:= \bigcup_{(n-2)j' + \iota' < (n-2)j + \iota} B_{.j' \iota'} \\ &= B_{.11} \cup B_{.12} \cup \dots \cup B_{.1(\iota-1)} \cup B_{.1\iota} \cup \dots \cup B_{.1(n-3)} \cup B_{.1(n-2)} \cup \\ &\quad B_{.21} \cup B_{.22} \cup \dots \cup B_{.2(\iota-1)} \cup B_{.2\iota} \cup \dots \cup B_{.2(n-3)} \cup B_{.2(n-2)} \cup \\ &\quad \dots \\ &\quad B_{.j1} \cup B_{.j2} \cup \dots \cup B_{.j(\iota-1)}. \end{aligned}$$

Next, we define the thresholds $\{\theta_{ij\iota}\}$ and $\theta_{1(m+1)}$. Denote

$$\omega_{j\iota} := ((j - 1)(n - 2) + (\iota - 1))W^3 + (n - 1)((j - 1)(n - 2) + (\iota - 1))W^2,$$

which is the expected number of neighbors of each $b_{ij\iota} \in B_{ij\iota}$ in $B_{\prec j \iota}$. For each fixed j , denote by i_j, i'_j the two indices such that $e_j = (A_{i_j}, A_{i'_j})$ with $i_j < i'_j$, and all $\theta_{ij\iota}$'s are defined as follow.

$$\begin{bmatrix} \theta_{1j1} & \theta_{2j1} & \dots & \theta_{nj1} \\ \theta_{1j2} & \theta_{2j2} & \dots & \theta_{nj2} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1jn} & \theta_{2jn} & \dots & \theta_{njn} \end{bmatrix} := \begin{bmatrix} \omega_{j1} + (1 - \Delta)W^2 & \omega_{j1} + (1 - \Delta)W^2 & \dots & \omega_{j1} + (1 - \Delta)W^2 \\ \omega_{j2} + (1 - \Delta)W^2 & \omega_{j2} + (1 - \Delta)W^2 & \dots & \omega_{j2} + (1 - \Delta)W^2 \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{jn} + (1 - \Delta)W^2 & \omega_{jn} + (1 - \Delta)W^2 & \dots & \omega_{jn} + (1 - \Delta)W^2 \end{bmatrix} +$$

$$\begin{bmatrix} & & & & \text{Column } i_j & & \text{Column } i'_j & & & & \\ & 1W^2 & & & 0 & \dots & 0 & \dots & (n - 3)W^2 & & (n - 2)W^2 \\ (n - 2)W^2 & & 1W^2 & & 0 & \dots & 0 & \dots & (n - 4)W^2 & & (n - 3)W^2 \\ (n - 3)W^2 & & (n - 2)W^2 & & 0 & \dots & 0 & \dots & (n - 5)W^2 & & (n - 4)W^2 \\ \vdots & & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 2W^2 & & 3W^2 & & 0 & \dots & 0 & \dots & (n - 2)W^2 & & 1W^2 \end{bmatrix}$$

Notice that for different $\iota_1, \iota_2 \in \{1, \dots, n - 2\}$, $(\theta_{1j\iota_1} - \omega_{j\iota_1}, \theta_{2j\iota_1} - \omega_{j\iota_1}, \dots, \theta_{nj\iota_1} - \omega_{j\iota_1})$ is a permutation of $(\omega_{1j\iota_2} - \omega_{j\iota_2}, \omega_{2j\iota_2} - \omega_{j\iota_2}, \dots, \omega_{nj\iota_2} - \omega_{j\iota_2})$. Specifically, for the second matrix above, excluding the i_j -th and the i'_j -th columns, the first row is an arithmetic progression $1W^2, 2W^2, (n - 2)W^2$, and the $(\iota + 1)$ -th row is obtained by cyclically shifting the ι -th row to the right by 1 unit.

Finally, for the threshold $\theta_{1(m+1)}$ of each vertex in the bundle C . We define

$$\theta_{1(m+1)} := m(n - 2)W^3 + (n - 1)m(n - 2)W^2 + (1 - \Delta)W^2.$$

As we will see later, $\theta_{1(m+1)}$ is slightly less than the expected number of neighbors of each $c \in C$ in $V \setminus C$, by an amount of $\Theta(\Delta W^2)$.

The High-level Ideas Before presenting rigorous arguments, we provide high level ideas of the reduction in this subsection.

We have constructed the hierarchy tree T , which corresponds to a graph distribution \mathcal{G} (refer to Definition 4). In the next subsection, we will show that a sample $G \sim \mathcal{G}$ can simulate the

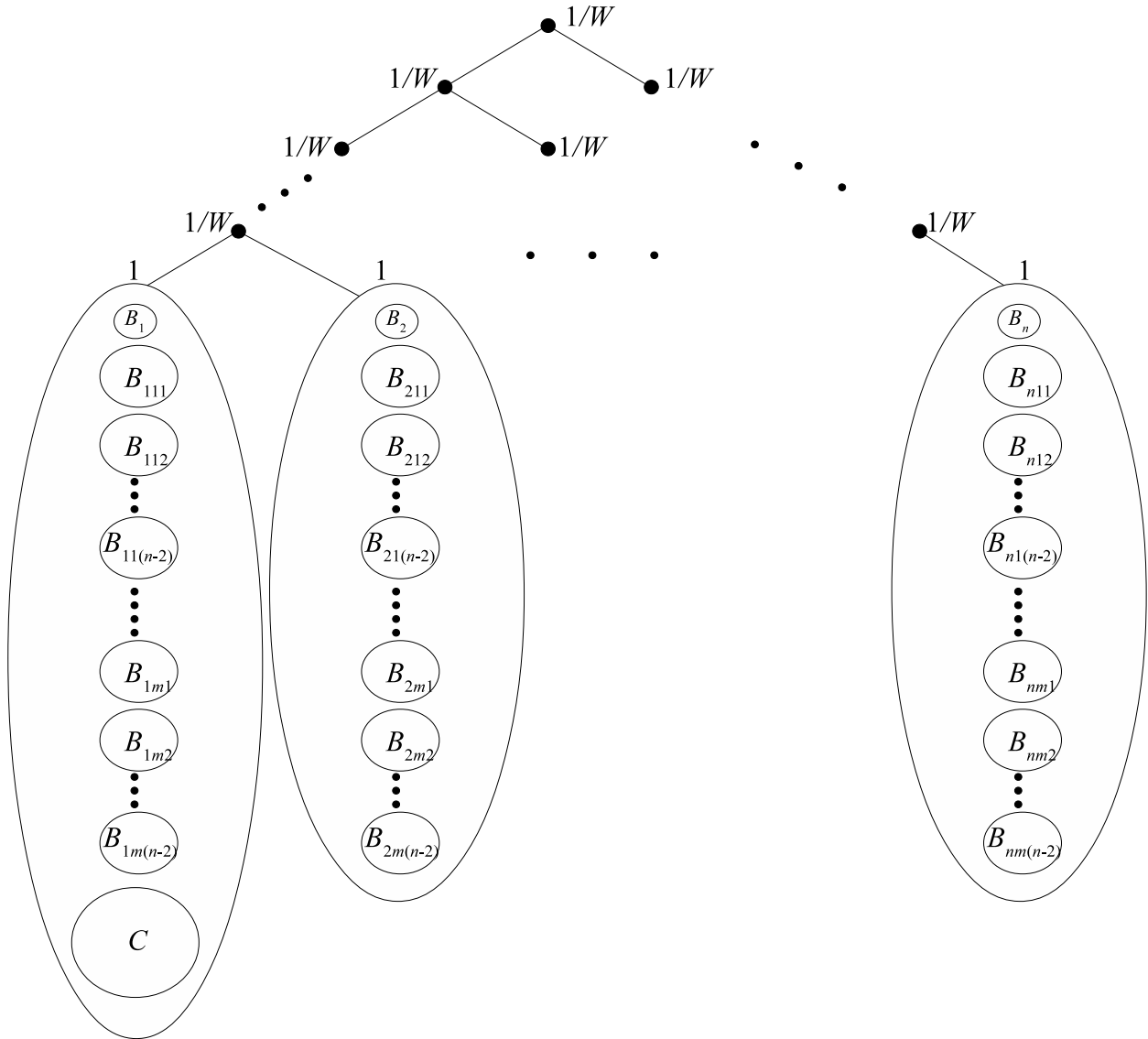


Figure 3: The construction of the hierarchy tree T .

corresponding VERTEXCOVER instance with high probability. In particular, we will say such samples are “good” samples, which we will define rigorously, and we will prove that a sample is good with probability $1 - o(1)$.

Given a VERTEXCOVER instance (\bar{G}, \bar{k}) , we consider the INFMAX instance (G, F, \mathcal{D}, k) , where G is a good sample, F, \mathcal{D} are as defined in Theorem 2 (or Theorem 3), and $k = \bar{k}W^2$.

Suppose we have a good sample G . If the VERTEXCOVER instance is a YES instance, we can find $\bar{S} \subseteq \bar{V}$ with $|\bar{S}| = \bar{k}$ such that \bar{S} covers all edges in \bar{E} . For each $A_i \in \bar{S}$, we choose W^2 seeds from the bundle B_i , so a total of $\bar{k}W^2 = k$ seeds are chosen.

Similar to what happens in Section 3, the cascade will flow level-by-level. In particular, for the first edge $e_1 \in \bar{E}$ and i_1, i'_1 such that $e_1 = (A_{i_1}, A_{i'_1})$, the vertices in the bundles B_{i_1} and $B_{i'_1}$ have the lowest threshold in the level $B_{.11}$. On the other hand, by our choice of k seeds, we have chosen W^2 seeds from one (or both) of B_i and $B_{i'}$. Calculations show that these seeds are just enough to infect all vertices in B_{i_1} and $B_{i'_1}$. The infection of these vertices will eventually infected the entire level $B_{.11}$, and similar analysis shows that the levels $B_{.12}, B_{.13}, \dots$ will be infected one-by-one. Finally, the cascade can reach the huge bundle C , and most vertices in G will be infected.

If the VERTEXCOVER instance is a NO instance, we can assume all seeds are chosen from $\{B_1, \dots, B_n\}$, as it is always a better idea to choose seeds from vertices having higher thresholds in a clique.¹⁰ We say that the i -th clique is activated if we have chosen almost W^2 seeds from B_i , or more than this number. We can draw an analogy between activating the i -th clique in INFMAX and picking the set A_i in VERTEXCOVER.

Since the VERTEXCOVER instance is a NO instance, certain element e_{j^*} is not covered, and we will show that the cascade will stop at one of the $n - 2$ levels $B_{.j^*1}, \dots, B_{.j^*(n-2)}$. Intuitively, the thresholds of vertices in these levels shift cyclically by our construction, and there exists a level whose vertices' thresholds are shifted to the position such that the cascade fails on all leaves. In particular, even if we put all $k = \bar{k}W^2$ seeds in a single bundle B_i , there exists a level ι such that $\theta_{i_{j^*\iota}}$ is large enough, making the cascade still fail on leaf i . On the other hand, there are only two leaves i_{j^*}, i'_{j^*} having lowest $\theta_{i_{j^*\iota}}$ in all levels $\iota = 1, \dots, n - 2$, which are exactly those i_{j^*}, i'_{j^*} with $e_{j^*} = (A_{i_{j^*}}, A_{i'_{j^*}})$. However, we have very few seeds (considerably fewer than W^2) on the i_{j^*} -th and the i'_{j^*} -th cliques, by our assumption that e_{j^*} is not covered.

Since the cascade will fail on a certain intermediate level, it cannot reach the huge bundle C . By making C contain most vertices in G (i.e., making M large enough), we can see that the number of infected vertices corresponding to a YES VERTEXCOVER instance is significantly higher, which implies both Theorem 2 and Theorem 3.

In the next two subsections, we will rigorously prove the correctness of our reduction.

Good Samplings In this subsection, we define “good” samplings $G \sim \mathcal{G}$ which are useful in the reduction from VERTEXCOVER, in the sense that G successfully simulates the VERTEXCOVER instance, and we show that a sample $G \sim \mathcal{G}$ is good with a high probability.

Firstly, consider a W^3 sized bundle $B_{ij\iota}$, and an arbitrary vertex v not in the i -th clique. Over all the samplings $G \sim \mathcal{G}$, v 's expected number of neighbors in $B_{ij\iota}$ is

$$\mathbb{E}_{G \sim \mathcal{G}} [|\Gamma(v) \cap B_{ij\iota}|] = \frac{1}{W} \cdot W^3 = W^2.$$

¹⁰Rigorously, this may not be true in the post-sampling case, where the seed-picker can see the sample G . The vertices not in $\{B_1, \dots, B_n\}$ may happen to have more neighbors across cliques, and the seed-picker can take advantage of this. We will reason about this later. However, for now, we assume all seeds are chosen from $\{B_1, \dots, B_n\}$.

Secondly, consider a set D_i of δW^2 vertices in the i -th clique, and a set D_{-i} of $(\bar{k}+1)W^2$ vertices that are not in the i -th clique, the expected total number of edges between D_i and D_{-i} is

$$\mathbb{E}_{G \sim \mathcal{G}} [|\{(u, v) : u \in D_i, v \in D_{-i}\}|] = \frac{1}{W} \cdot \delta W^2 \cdot (\bar{k}+1)W^2 = \delta(\bar{k}+1)W^3.$$

We define a sampling $G \sim \mathcal{G}$ to be “good” if the above two numbers roughly concentrate on their expectations.

Definition 12. *A sampling $G \sim \mathcal{G}$ is **good** if the following holds.*

1. For all $i = 1, \dots, n$, $j = 1, \dots, m$ and $\iota = 1, \dots, n-2$, and any vertex v not in the i -th clique,

$$(1 - \delta)W^2 < |\Gamma(v) \cap B_{ij\iota}| < (1 + \delta)W^2.$$

2. For any set D_i of δW^2 vertices in the i -th clique, and any set D_{-i} of $(\bar{k}+1)W^2$ vertices that are not in the i -th clique, the number of edges between D_i and D_{-i} is less than $W^{3.6}$:

$$|\{(u, v) : u \in D_i, v \in D_{-i}\}| < W^{3.6}.$$

The following lemma shows that a sampling $G \sim \mathcal{G}$ is good with high probability.

Lemma 1. *A sampling $G \sim \mathcal{G}$ is good with probability more than $1 - e^{-\sqrt{W}}$.*

Proof. We apply Chernoff-Hoeffding inequality and union bounds to show this lemma. In a random sample $G \sim \mathcal{G}$, for each $i = 1, \dots, n$; $j = 1, \dots, m$; $\iota = 1, \dots, n-2$ and v , requirement 1 in Definition 12 fails with probability

$$\Pr [|W^2 - |\Gamma(v) \cap B_{ij\iota}| | \geq \delta W^2] \leq 2 \exp \left(-\frac{1}{2} (\delta W^2)^2 \frac{1}{W^3} \right) < e^{-W^{0.6}},$$

where the last inequality is due to $(\delta W^2)^2 = \frac{1}{\bar{k}^2} m^{38} n^{36} > W^{3.6}$.

For each D_i and D_{-i} , requirement 2 in Definition 12 fails with probability

$$\Pr [|\{(u, v) : u \in D_i, v \in D_{-i}\}| \geq W^{3.6}] \leq \exp \left(-\frac{1}{2} \frac{(W^{3.6} - \delta(\bar{k}+1)W^3)^2}{\delta W^2 \cdot (\bar{k}+1)W^2} \right) < e^{-W^3}.$$

By a union bound, the probability that a sample $G \sim \mathcal{G}$ is not good is

$$\begin{aligned} \Pr[\text{not good}] &< nm(n-2)N e^{-W^{0.6}} + \binom{N}{\delta W^2} \binom{N}{\bar{k}W^2} e^{-W^3} \\ &< N^2 e^{-W^{0.6}} + N^{\delta W^2 + \bar{k}W^2} e^{-W^3} \\ &= e^{2 \log N} e^{-W^{0.6}} + e^{(\delta W^2 + \bar{k}W^2) \log N} e^{-W^3} \\ &< e^{-\sqrt{W}}, \quad (\text{as } N = \text{poly}(W), \text{ which implies } \log N = o(W^c) \text{ for arbitrary } c > 0) \end{aligned}$$

which immediately implies the lemma. \square

The Reduction Correctness In this section, we show that INFMAX on a good sample $G \sim \mathcal{G}$ simulates the VERTEXCOVER problem.

Lemma 2. *Consider INFMAX with $k = \bar{k}W^2$ seeds. For any good sample $G \sim \mathcal{G}$,*

1. *if the VERTEXCOVER instance is a YES instance, a total of $\bar{k}W^2 + nm(n-2)W^3 + M$ vertices can be infected by properly choosing the k seeds;*
2. *if the VERTEXCOVER instance is a NO instance, at most $N - M$ vertices can be infected for any choices of the k seeds.*

Proof of 1. Suppose the VERTEXCOVER instance is a YES instance. Let \bar{S} be the choice of \bar{k} vertices in VERTEXCOVER instance that covers all edges in \bar{E} . As mentioned earlier, we can assume $A_1 \in \bar{S}$. For each $A_i \in \bar{S}$, we choose W^2 seeds from the bundle B_{i_1} , so a total of $\bar{k}W^2 = k$ seeds are chosen.

We show that all vertices in the level $B_{\cdot 11}$ will be infected. Consider $e_1 = (A_{i_1}, A_{i'_1})$ with $i_1 < i'_1$. By the fact the VERTEXCOVER instance is a YES instance and the way we choose the seeds, W^2 vertices in either B_{i_1} or $B_{i'_1}$, or both, are seeded. Assume without loss of generality that W^2 vertices from B_{i_1} are seeded, then all the vertices in the bundle $B_{i_1 11}$, having threshold $\theta_{i_1 11} = \omega_{11} + (1 - \Delta)W^2 + 0 = (1 - \Delta)W^2 < W^2$ will be infected. As for the vertices in $B_{i'_1 11}$, they will be infected in the same way if W^2 vertices from $B_{i'_1}$ are also seeded. On the other hand, if no vertex in $B_{i'_1}$ is seeded, all vertices in $B_{i'_1 11}$ will be infected due to the influence of $B_{i_1 11}$. This is because 1) each vertex in $B_{i'_1 11}$ has more than $(1 - \delta)W^2$ infected neighbors in $B_{i_1 11}$ by requirement 1 of Definition 12, and 2) each vertex in $B_{i'_1 11}$ has threshold $(1 - \Delta)W^2 < (1 - \delta)W^2$. In the next $n - 2$ iterations, by a careful calculation and based on requirement 1 of Definition 12, all vertices in the remaining $n - 2$ bundles $\{B_{i 11}\}_{i \neq i_1, i'_1}$ will be infected in the following order:

$$B_{111} \rightarrow B_{211} \rightarrow \cdots B_{(i_1-1)11} \rightarrow B_{(i_1+1)11} \rightarrow \cdots B_{(i'_1-1)11} \rightarrow B_{(i'_1+1)11} \rightarrow \cdots \rightarrow B_{n11}. \quad (1)$$

Therefore, the entire level $B_{\cdot 11}$ will be infected.

By similar analyses, we will show that the next level $B_{\cdot 12}$ will be infected after the previous level $B_{\cdot 11}$. Again, assume without loss of generality that W^2 seeds in B_{i_1} are chosen. (Remember that the first $n-2$ levels are for edge $e_1 \in \bar{E}$, so we are still working on e_1 .) Each vertex in $B_{i_1 12}$ has $(W^2 + W^3)$ infected neighbor in the i_1 -th clique, and has more than $(n-1)(1-\delta)W^2$ infected neighbors in $\{B_{i 11}\}_{i \neq i_1}$, which is a total of more than $W^3 + nW^2 - (n-1)\delta W^2$ neighbors. Moreover, each vertex in $B_{i_1 12}$ has threshold $\theta_{i_1 12} = \omega_{12} + (1 - \Delta)W^2 + 0 = W^3 + (n-1)W^2 + (1 - \Delta)W^2 = W^3 + nW^2 - \Delta W^2$ which is less than the number of infected neighbors, as $-\Delta < -(n-1)\delta$. Therefore, all vertices in $B_{i_1 12}$ will be infected. As for the vertices in $B_{i'_1 12}$, following the analysis in the last paragraph, they will be infected at the same iteration if W^2 vertices in $B_{i'_1}$ are seeded, and they will be infected at the next iteration due to the extra influence from $B_{i_1 12}$ if not. Finally, the remaining $n - 2$ bundles $\{B_{i 12}\}_{i \neq i_1, i'_1}$ will be infected in the following order:

$$B_{212} \rightarrow B_{312} \rightarrow \cdots B_{(i_1-1)12} \rightarrow B_{(i_1+1)12} \rightarrow \cdots B_{(i'_1-1)12} \rightarrow B_{(i'_1+1)12} \rightarrow \cdots \rightarrow B_{n12} \rightarrow B_{112}, \quad (2)$$

which is similar to (1), but is cyclically shifted to the left by 1 unit, due to our cyclic construction of the thresholds. Thus, we have shown that the level $B_{\cdot 12}$ will be infected after the previous level $B_{\cdot 11}$.

Following the same analyses, we can conclude that all levels will be infected in the following order:

$$B_{\cdot 11} \rightarrow B_{\cdot 12} \rightarrow \cdots \rightarrow B_{\cdot 1(n-2)} \rightarrow$$

$$\begin{aligned}
& B_{.21} \rightarrow B_{.22} \rightarrow \cdots \rightarrow B_{.2(n-2)} \rightarrow \\
& \qquad \qquad \qquad \cdots \\
& B_{.m1} \rightarrow B_{.m2} \rightarrow \cdots \rightarrow B_{.m(n-2)}.
\end{aligned}$$

Lastly, each vertex $c \in C$ has $W^2 + m(n-2)W^3$ infected neighbors in the 1-st clique (notice that we have assume $A_1 \in \bar{S}$, which implies W^2 vertices in B_1 are seeded, which contributes W^2 infected neighbors), and more than $(n-1) \cdot m(n-2)(1-\delta)W^2$ infected neighbors from the other $n-1$ cliques, which is a total of $m(n-2)W^3 + (n-1)m(n-2)W^2 + W^2 - (n-1)m(n-2)\delta W^2$ neighbors. In addition, c has threshold $\theta_{1(m+1)} = m(n-2)W^3 + (n-1)m(n-2)W^2 + (1-\Delta)W^2$, which is less than the number of infected neighbors, as we have $-\Delta = -mn^2\delta < -(n-1)m(n-2)\delta$. Consequently, all vertices in C will be infected. By summing up the total number of infected vertices, we conclude the first part of this lemma. \square

Proof of 2. Suppose the VERTEXCOVER instance is a NO instance. For those $n\bar{k}W^2$ vertices in $\{B_i\}_{i=1,\dots,n}$ having threshold ∞ , they will not be infected unless being seeded, which means at least $(n-1)\bar{k}W^2$ of them will not be infected. To show that the total number of infected vertices cannot exceed $N - M$, it is enough to show that at most $(n-1)\bar{k}W^2$ vertices can be infected in the bundle C of M vertices. We will show the following stronger claim.

Proposition 1. *If the VERTEXCOVER instance is a NO instance, all vertices in C will not be infected unless being seeded.*

To show Proposition 1, we show that the cascade will stop at an intermediate level. We will first identify this level, and then show this claim in Proposition 2.

Consider an arbitrary seed set S (with $|S| = k$). Let S_i be the seeds chosen from the i -th clique, and $k_i = |S_i|$ so that $\sum_{i=1}^n k_i = k$. We say that the i -th clique is activated if $k_i \geq (1 - 9\Delta)W^2$. Since $(\bar{k} + 1)(1 - 9\Delta)W^2 = \bar{k}W^2 + \left(1 - \frac{9}{10} - \frac{9}{10\bar{k}}\right)W^2 > k$, at most \bar{k} cliques can be activated.

If we draw an analogy between activating a clique and picking a subset in VERTEXCOVER, by the fact that the VERTEXCOVER instance is a NO instance, there exists j^* where $e_{j^*} = (A_{i_{j^*}}, A_{i'_{j^*}})$ such that both i_{j^*} -th and i'_{j^*} -th cliques are not activated. For the ease of illustration, assume without loss of generality that $i_{j^*} = n-1$ and $i'_{j^*} = n$. Since we have assumed $n > \bar{k}^2 + 2$, there exists $i^* \leq n-1-\bar{k}$ such that the i^* -th, the (i^*+1) -th, ..., and the $(i^*+\bar{k}-1)$ -th cliques are not activated. (If we have an activated clique within any \bar{k} consecutive cliques in the first $n-2$ cliques, the total number of activated cliques is at least $\frac{n-2}{\bar{k}} > \bar{k}$, which is a contradiction.) We will show that the cascade stops at the level $B_{.j^*i^*}$. That is, there are only $o(W^3)$ infected vertices in

$$\left(\bigcup_{(n-2)j+i \geq (n-2)j^*+i^*} B_{.ji} \right) \cup C = V \setminus (B_1 \cup \cdots \cup B_n \cup B_{\prec j^*i^*}).$$

We will show that this is true even in the case that all vertices in the previous $(n-2)(j^*-1)+i^*-1$ levels (i.e., those in $B_{\prec j^*i^*}$) are infected.

Proposition 2. *There are only $o(W^3)$ infected vertices in the level $B_{.j^*i^*}$, given that all vertices in $B_{\prec j^*i^*}$ and at most $\bar{k}W^2$ vertices elsewhere (i.e., in $V \setminus B_{\prec j^*i^*}$) are infected.*

The “ $\bar{k}W^2$ vertices elsewhere” mentioned in Proposition 2 refer to the $k = \bar{k}W^2$ seeds. Notice that the seed-picker may choose the seeds outside $B_{\prec j^*i^*}$, and Proposition 2 holds even if all vertices in $B_{\prec j^*i^*}$ are infected and the k seeds are all outside $B_{\prec j^*i^*}$.

Before proving Proposition 2, we remark that Proposition 2 immediately implies Proposition 1: the vertices in the later levels $B_{.j^*(i^*+1)}, B_{.j^*(i^*+2)}, \dots, B_{.m(n-2)}$ have thresholds even higher than the thresholds of vertices in $B_{.j^*i^*}$, and the thresholds increase by $\Theta(W^3)$ for each next level.

Proof of Proposition 2. Suppose all vertices in $B_{\prec j^*i^*}$ and at most $\bar{k}W^2$ vertices elsewhere are infected after certain cascade iteration t . We will first show that less than δW^2 not-seeded vertices can be infected in each bundle $B_{ij^*i^*}$ for $i = 1, \dots, n$ in the next cascade iteration $t + 1$. Then, we will show the same claim for later iterations.

For each vertex in the bundle $B_{nj^*i^*}$, by requirement 1 of Definition 12, the number of infected neighbors among the vertices in $B_{\prec j^*i^*}$ is less than

$$\underbrace{((n-2)(j^*-1) + i^* - 1)W^3}_{\text{from the } n\text{-th clique}} + \underbrace{(n-1) \cdot ((n-2)(j^*-1) + i^* - 1) \cdot (1 + \delta)W^2}_{\text{from the other } n-1 \text{ cliques}} < \omega_{j^*i^*} + \Delta W^2. \quad (3)$$

For each vertex in the bundle $B_{nj^*i^*}$, we have already counted the number of infected neighbors in $B_{\prec j^*i^*}$. Next, we consider the infected neighbors in $V \setminus B_{\prec j^*i^*}$. There are at most $\bar{k}W^2$ of them by our assumption, and they are the seeds $S = \bigcup_{i=1}^n S_n$.

The number of infected neighbors among seed set S_n contributes at most $k_n < (1 - 9\Delta)W^2$, as we have assumed the n -th clique is not activated. Summing up this and (3), the total number of infected neighbors in $B_{\prec j^*i^*} \cup S_n$ is at most $\omega_{j^*i^*} + (1 - 8\Delta)W^2$. Since by our construction $\theta_{nj^*i^*} = \omega_{j^*i^*} + (1 - \Delta)W^2 + 0$, in order to have δW^2 not-seeded vertices infected, the number of edges between each of these δW^2 vertices and $\bigcup_{i=1}^{n-1} S_i$ should be more than

$$7\Delta W^2, 7\Delta W^2 - 1, 7\Delta W^2 - 2, \dots, 7\Delta W^2 - \delta W^2 + 1$$

respectively. This requires a total of

$$\sum_{t=0}^{\delta W^2 - 1} (7\Delta W^2 - t) > \delta W^2 (7\Delta W^2 - \delta W^2 + 1) > W^{3.6}$$

edges, where the last inequality is based on the fact $\delta W^2 = \frac{1}{10\bar{k}} m^{19} n^{18} \gg W^{1.6}$. Since $\sum_{i=1}^{n-1} k_i < (\bar{k} + 1)W^2$, this is a contradiction to requirement 2 of Definition 12.

For exactly the same reason, we can only have less than δW^2 not-seeded vertices infected in the bundle $B_{(n-1)j^*i^*}$, as $\theta_{(n-1)j^*i^*} = \theta_{nj^*i^*}$.

Next, we consider these \bar{k} bundles: $B_{i^*j^*i^*}, B_{(i^*+1)j^*i^*}, \dots, B_{(i^*+\bar{k}-1)j^*i^*}$, whose corresponding cliques $i^*, i^* + 1, \dots, i^* + \bar{k} - 1$ are not activated by our assumption. Based on our construction, the vertices in these bundles have thresholds

$$\omega_{j^*i^*} + (1 - \Delta)W^2 + 1W^2, \omega_{j^*i^*} + (1 - \Delta)W^2 + 2W^2, \dots, \omega_{j^*i^*} + (1 - \Delta)W^2 + \bar{k}W^2$$

respectively, which are all more than $\theta_{nj^*i^*}$. By the same arguments, we can show that having δW^2 not-seeded vertices infected in any of these bundle requires even more edges, which contradicts to requirement 2 of Definition 12.

For each of the remaining $n - 2 - \bar{k}$ bundles $B_{ij^*i^*}$ with $i \neq i, i + 1, \dots, i + \bar{k} - 1, n - 1, n$, although the corresponding i -th clique may be activated, but the threshold $\theta_{ij^*i^*}$ is at least $\omega_{j^*i^*} + (1 - \Delta)W^2 + (\bar{k} + 1)W^2$. The number of seeds chosen in the i -th clique k_i cannot offset the term $(\bar{k} + 1)W^2$. Therefore, applying the same arguments will still show us that less than δW^2 not-seeded vertices can be infected in each of these bundles.

We have shown that less than δW^2 not-seeded vertices can be infected in each bundle $B_{ij^*i^*}$ in iteration $t + 1$. To show this claim for future iterations, assume for the sake of contradiction that 1) at iteration $t^* > t + 1$, less than δW^2 not-seeded vertices are infected in each bundle $B_{ij^*i^*}$, and 2) at iteration $t^* + 1$, for certain i^* we have at least δW^2 not-seeded vertices infected in the bundle $B_{i^*j^*i^*}$. Denote by D_{-i^*} the set of those vertices outside the i^* -th clique which are infected during the iterations $t + 1, t + 2, \dots, t^*$, and D_{i^*} be the set of those vertices in the i^* -th clique which are infected during the iterations $t + 1, t + 2, \dots, t^*, t^* + 1$. Following the same arguments, for some δW^2 vertices from D_{i^*} , the number of edges between each of these δW^2 vertices and $D_{-i^*} \cup S$ should be more than

$$7\Delta W^2, 7\Delta W^2 - 1, 7\Delta W^2 - 2, \dots, 7\Delta W^2 - \delta W^2 + 1$$

respectively, whose summation is more than $W^{3.6}$. On the other hand, since $|D_{-i^*}| < (n-1) \cdot \delta W^2 < W^2$, we have $|D_{-i^*} \cup S| < (1 + \bar{k})W^2$, which again contradicts to requirement 2 of Definition 12. Therefore, we conclude Proposition 2. \square

As we have remarked that Proposition 2 implies Proposition 1, we conclude the second part of Lemma 2. \square

Finally, by making M sufficiently large, both Theorem 2 and Theorem 3 follow from Lemma 1 and Lemma 2.

5 Hierarchical Blockmodel with One-Way Influence

In this section, we consider a variant to the hierarchical blockmodel in which the influence between any two vertex-blocks can only be “one-way”. To each node in the hierarchy tree, a *sign* is assigned deciding the directions of the edges between the two vertex-blocks associated to its two children. For example, let t be a node in the hierarchy tree, and t_L, t_R be its left child and right child respectively. If t has a positive sign, then all edges between $V(t_L)$ and $V(t_R)$ are from $V(t_L)$ to $V(t_R)$; otherwise, these edges are from $V(t_R)$ to $V(t_L)$. In this manner, the influence between $V(t_L)$ and $V(t_R)$ is one-way.

In INFMAX, the seed-picker needs to decide not only the choice of those k seeds, but also the sign at each tree node. That is, the algorithm to INFMAX problem should also output the optimal directions of influence between each pair of vertex-blocks.

5.1 A Dynamic Program Algorithm

We present a dynamic program based algorithm for INFMAX for this variant of the hierarchical blockmodel, when the thresholds of the vertices are deterministic. Our algorithm makes use of the following observation: *for a tree node t , the influence of vertex-block $V(t)$ to $V \setminus V(t)$ only depends on the **number** of infected vertices in $V(t)$.*

For each tree node $t \in V_T$, each $i = 1, \dots, k$, and each $\nu = 0, 1, \dots, |V_T|$, define $H[t, i, \nu]$ be the smallest positive real number γ satisfying the following:

- given that the threshold of each vertex is updated to $\theta_v \leftarrow \theta_v - \gamma$, where we assume the vertex with $\theta_v - \gamma \leq 0$ is infected immediately, we can choose i seeds in $V(t)$ such that at least ν vertices in $V(t)$ will be infected (due to the influence of these i seeds).

Intuitively, this means we can infected ν vertices by i seeds, given that the influence from infected vertices outside $V(t)$ is $H[t, i, \nu]$.

If t is a leaf, the subgraph induced by $V(t)$ is a clique in which all the $|V(t)|(|V(t)| - 1)$ edges have the equal weight. Obviously, the optimal strategy is to place the i seeds on those vertices with highest thresholds. We proposes Algorithm 1 to calculate $H[t, i, \nu]$ for each leaf t .

Algorithm 1: Initialization for An Leaf t

Input: vertex set $V(t)$, weight of each edge $w(t)$, threshold set $\{\theta_v\}_{v \in V(t)}$, integers i, ν
Output: $H[t, i, \nu]$ for leaf t
for each vertex $v \in V(t)$ **do**
 | update $\theta_v \leftarrow \theta_v - i \cdot w(t)$
end
if $\nu \leq |\{\theta_v : \theta_v \leq 0\}| + i$ **then**
 | set $H[t, i, \nu] = 0$
else
 | set $H[t, i, \nu]$ be the $(\nu - i)$ -th smallest threshold in $\{\theta_v\}_{v \in V(t)}$
end
return $H[t, i, \nu]$

If t is not a leaf, we aim to find a recurrence between $H[t, i, \nu]$ and $H[t_L, i_L, \nu_L], H[t_R, i_R, \nu_R]$. Suppose the sign of t is positive, and there are ν_L infected vertices in $V(t_L)$. Their influence to $V(t_R)$ is $\nu_L \cdot w(t)$ where $w(t)$ is the weight of t reflecting the weight of all edges from $V(t_L)$ to $V(t_R)$. We have the similar observation in the case that the sign of t is negative.

By considering all decompositions $i = i_L + i_R$ and $\nu = \nu_L + \nu_R$, if the sign of t is positive, we have

$$H^+[t, i, \nu] = \min_{i_L=0, \dots, i; \nu_L=0, \dots, \nu} \left\{ \max (H[t_L, i_L, \nu_L], H[t_R, i - i_L, \nu - \nu_L] - \nu_L \cdot w(t)) \right\};$$

if the sign of t is negative, we have

$$H^-[t, i, \nu] = \min_{i_R=0, \dots, i; \nu_R=0, \dots, \nu} \left\{ \max (H[t_L, i - i_R, \nu - \nu_R] - \nu_R \cdot w(t), H[t_R, i_R, \nu_R]) \right\},$$

where we set $H[t, i, \nu] = \infty$ if $\nu > |V(t)|$. Finally, we decide the sign of t :

$$H[t, i, \nu] = \min (H^+[t, i, \nu], H^-[t, i, \nu]).$$

Define the *height* of $t \in V_T$ be the length of the path to t 's deepest descendent. The following Algorithm 2 solves INFMAX for the hierarchical blockmodel with one-way influence. It is straightforward to check that Algorithm 2 runs in time $O(N^3 k^2)$.

5.2 Further Discussions

We have seen inapproximability results in Section 3 and Section 4 for INFMAX on the (stochastic) hierarchical blockmodel. Our algorithm in this section reveals the intrinsic reason why these problems are difficult.

In the hard INFMAX instances in Figure 2 and Figure 3, we constructed the hierarchy tree by creating n branches corresponding to the n vertices in VERTEXCOVER. In the case the VERTEXCOVER instance is a YES instance, the influence of the properly chosen seeds actually passes through

Algorithm 2: Dynamic Programming Algorithm for Hierarchical Blockmodel INFMAX with One-Way Influence

Input: hierarchical blockmodel $G = (V, T)$, threshold set $\{\theta_v\}_{v \in V}$, integer k
Output: $S \subseteq V$ such that $|S| = k$ and S maximizes $\sigma(S)$
for each height $i = 0, 1, \dots, h$ **do**
 for each node $t \in V_T$ with height i **do**
 if t is a leaf **then**
 initialize $H[t, i, \nu]$ by Algorithm 1 for all $i = 0, 1, \dots, k$ and $\nu = 0, 1, \dots, N$
 else
 for each $i = 0, 1, \dots, k$ and $\nu = 0, 1, \dots, N$ **do**
 $H^+[t, i, \nu] =$
 $\min_{i_L=0, \dots, i; \nu_L=0, \dots, \nu} \left\{ \max(H[t_L, i_L, \nu_L], H[t_R, i - i_L, \nu - \nu_L] - \nu_L \cdot w(t)) \right\}$
 $H^-[t, i, \nu] =$
 $\min_{i_R=0, \dots, i; \nu_R=0, \dots, \nu} \left\{ \max(H[t_L, i - i_R, \nu - \nu_R] - \nu_R \cdot w(t), H[t_R, i_R, \nu_R]) \right\}$
 $H[t, i, \nu] = \min(H^+[t, i, \nu], H^-[t, i, \nu])$
 end
 end
 end
end
return the maximum ν such that $H[r, k, \nu] = 0$, for root node r

these n branches “back-and-forth” frequently: the infected vertices in branch A_i make vertices in branch A_j infected, while these newly infected vertices in A_j may have backward influence to A_i , and cause more infected vertices in A_i . The cascade process forms a back-and-forth “path” among these branches. In order to optimally choose the seeds, the seed-picker needs to evaluate all the possible paths of cascade and decide the best one. It is exactly the huge number of such possible paths makes INFMAX hard.

On the other hand, if there is no such bidirectional effect, INFMAX becomes easy on the hierarchical blockmodel, as our algorithm in this section suggests.

As mentioned in the related work section, Angell and Schoenebeck [1] show that a generalization of this algorithm works well empirically. This perhaps indicates that the bidirectional influence is, in average case, not often so important in realistic settings.

6 2-Submodular Influence Maximization

We prove the following theorem in this section.

Theorem 4. *For any fixed 2-submodular f , there exists a constant τ depending on f such that INFMAX with universal local influence model I_G^f is NP-hard to approximate to within factor N^τ .*

The sequence notation $(a_i)_{i=0,1,2,\dots}$ is used to represent f in this section. Because f is 2-submodular, we have $a_0 = 0$ and $a_2 > 2a_1$. We denote $p^* = \lim_{i \rightarrow \infty} a_i$, which exists because (a_i) is increasing and bounded. We consider two cases: $a_1 > 0$ and $a_1 = 0$. We note that we have $a_2 > 0$ by the 2-submodular assumption. In the case $a_1 > 0$, we will first assume the graph is directed, and later we will show that this assumption is not essential.

Section 6.1 to Section 6.6 discuss the case $a_1 > 0$, while Section 6.7 discusses the case $a_1 = 0$.

6.1 Proof of Theorem 4 for $a_1 > 0$ with Directed Graphs

We first define the following AND gadget which simulates the logical AND operation. The construction of this AND gadget is deferred to Section 6.3–6.5. We note that the nonsubmodularity property $a_2 > 2a_1$ plays an important role in the construction of the AND gadget. In particular, the construction of AND gadget uses a smaller gadget called “probability filter gadget” as a building block (see Figure 5), and 2-submodularity is essential for constructing the probability filter gadget (refer to Section 6.3.2 for details).

Definition 13. *An $(I, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget takes I sets which each contains Λ vertices as input, and outputs one vertex such that*

1. *if all the vertices in all I sets are infected independently with probabilities less than $\frac{11}{10}p_0$, and moreover the infection probabilities of the vertices in at least one input set are less than $\frac{1}{2}p_0$, then the output vertex will be infected with probability less than ε_1 ;*
2. *if all the vertices in all I sets are infected independently with probabilities in the interval $(p_0, \frac{11}{10}p_0)$, the output vertex will be infected with probability in $(p_2 - \varepsilon_2, p_2]$,*

We remark that the choices for both factors of p_0 in 1 of the above definition, $\frac{11}{10}$ and $\frac{1}{2}$, are only required to be close enough to 1 and 0 respectively. We aim to simulate the case where at least one of the inputs is not “active” (being far from the threshold p_0) and the other ones are not “too active” (being at most somewhere around the threshold p_0), in which case the AND gadget outputs “false” (such that the output vertex is infected with negligible probability ε_1).

With the choice of the seven parameters satisfying the relation in the below lemma, we can construct the AND gadget.

Lemma 3. *Given any 2-submodular function f with $a_1 > 0$, any constant threshold $p_0 > 0$ and any $I = 2^\ell$ that is an integer power of 2, there exists a constant $p_2 > 0$ depending on p_0 and f such that for any $\varepsilon_1 > 0$ and any constant $\varepsilon_2 > 0$, we can construct an $(I, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget with $\Lambda = O((1/\varepsilon_1)^{c_1} I^{c_2})$, and the numbers of vertices and edges in this AND gadget are both $O((1/\varepsilon_1)^{c_1} I^{c_2+1})$, where c_1 and c_2 are two constants.*

The following lemma is needed in the next section for the proof of Theorem 4 for undirected graphs.

Lemma 4. *Given any 2-submodular function f with $a_1 > 0$ and any $I = 2^\ell$ that is an integer power of 2, there exists $p_2 > 0$ such that for any $\varepsilon_1 > 0$ and any constant $\varepsilon_2 > 0$, we can construct an $(I, \Lambda, p^*(p_2 - \varepsilon_2), p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget. We have $\Lambda = O((1/\varepsilon_1)^{c_1} I^{c_2})$ and the AND gadget contains $O((1/\varepsilon_1)^{c_1} I^{c_2+1})$ vertices and $O((1/\varepsilon_1)^{c_1} I^{c_2+1})$ edges, where c_1 and c_2 are two constants.*

Notice that Lemma 3 does not imply Lemma 4: in Lemma 3, we first fix the third parameter p_0 , and the existence of the fourth parameter p_2 relies on the third; in Lemma 4, we simultaneously fix the third and the fourth parameters.

The construction of the AND gadget and the proof of Lemma 3 and Lemma 4 are deferred to Section 6.5. In this section, we aim to prove Theorem 4 for $a_1 > 0$ with directed graphs assuming Lemma 3, while we do not need Lemma 4 at this moment. We remark that the construction of AND gadget requires no directed edges, although we consider directed graph in this section.

6.1.1 A Reduction from SETCOVER

We prove the theorem by a reduction from SETCOVER.

Definition 14. *Given a universe U of n elements, a set of κ subsets $A = \{A_i \mid A_i \subseteq U\}$, and a positive integer k , the SETCOVER problem asks if we can choose k subsets $\{A_{i_1}, \dots, A_{i_k}\} \subseteq A$ such that $A_{i_1} \cup \dots \cup A_{i_k} = U$.*

Without loss of generality, we will assume $\kappa = O(n)$.¹¹ We will also assume that each element in U is covered by at least one subset A_i in SETCOVER (otherwise we know for sure the instance is a NO instance). In addition, we assume the number of elements $n = |U|$ is a integer power of 2, as we can add elements into U and let these elements be included in all sets A_i in the case n is not an integer power of 2.

We construct a graph G with N vertices which consists of two parts: the set cover part and the verification part, where the set cover part simulates the SETCOVER instance and the verification part verifies if all the elements in the SETCOVER instance are covered. The construction is shown in Figure 4.

Define $\varepsilon = 2(p^* - a_{\lfloor a_1 n \rfloor})$ which approaches to 0 as $n \rightarrow \infty$ if $a_1 > 0$. According to Lemma 3, for $p_0 = a_1(p^* - \varepsilon)$ and $I = n$, there exists a constant $p_2 > 0$, such that if we set $\varepsilon_1 = \frac{1}{n}$ and $\varepsilon_2 = \frac{1}{100}p_2$, we can construct an $(n, \Lambda, p^* - \varepsilon, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget, where $\Lambda = O((1/\varepsilon_1)^{c_1} n^{c_2}) = O(n^{c_1+c_2})$. We will use this AND gadget later. Define $M_1 = n^{c_1+c_2+10}$, $M_2 = n^2$, and $m = M_2\Lambda$.

The Set Cover Part Given a SETCOVER instance, we use a single vertex to represent a subset A_i and a clique of size m to represent each element in U . If an element is in a subset, we create m directed edges from the vertex representing the subset to each the m vertices in the clique representing the element.

The Verification Part We construct the $(n, \Lambda, a_1(p^* - \varepsilon), p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget mentioned. We associate each of the n cliques to one of the n inputs of this AND gadget, such that a matching is form between the n cliques and the n inputs. For each of the n cliques and its associated input, we choose Λ vertices from the clique, and connect them to the Λ vertices of the associated input by Λ directed edges. We create M_1 vertices and let the output vertex v of the AND gadget be connected to these M_1 vertices with undirected edges. Then, we duplicate the AND gadget and the attached M_1 vertices to a total of M_2 copies such that the vertices at the input ends of the AND gadgets in all these M_2 copies are connected from the different vertices in the n cliques as inputs. This, in particular, justifies our choice of clique size $m = M_2\Lambda$.

The Size of the Construction To show that the reduction is in polynomial time, it is enough to show that the number of vertices N in the graph G we constructed is a polynomial of n . According to Lemma 3, the AND gadget has $O((1/\varepsilon_1)^{c_1} n^{c_2+1}) = O(n^{c_1+c_2+1})$ vertices. We have

$$N = \kappa + mn + M_2 (O(n^{c_1+c_2+1}) + M_1) = \kappa + mn + \Theta(n^{c_1+c_2+12}) = \Theta(n^{c_1+c_2+12}),$$

where $\kappa + mn$ is the size for the set cover part and $M_2 (O(n^{c_1+c_2+1}) + M_1)$ is the size for the verification part.

¹¹One way to justify this assumption is to consider VERTEXCOVER, which can be viewed as a special case of SETCOVER by viewing vertices as subsets and edges as elements. In a connected graph, the number of vertices κ never exceeds $O(n)$, if n is the number of edges.

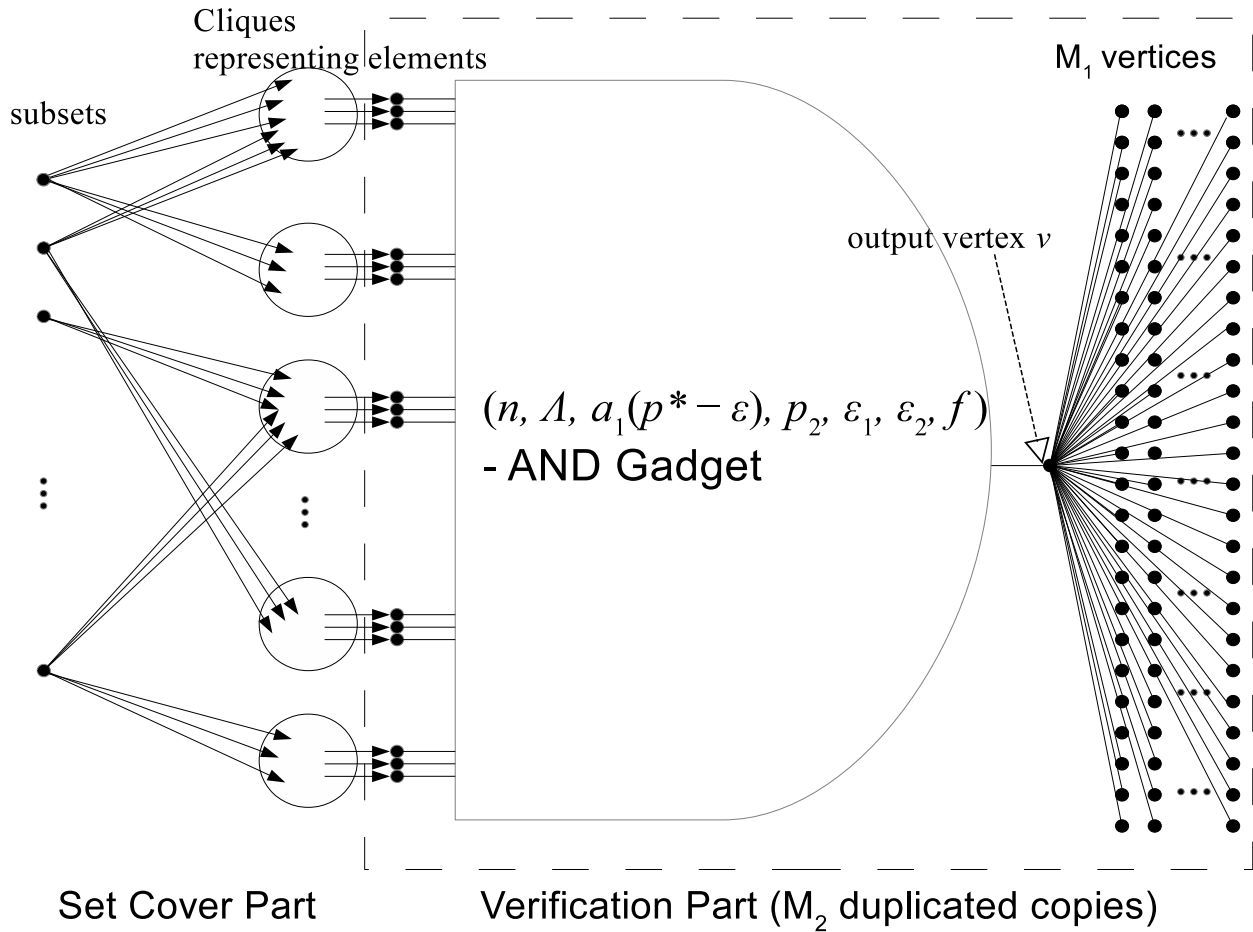


Figure 4: The high-level structure of the reduction

Finally, the lemma below immediately concludes Theorem 4 for the case $a_1 > 0$ with directed edges.

Lemma 5. *If the SETCOVER instance is a YES instance, by choosing k seeds appropriately, we can infect at least $\Theta(n^{c_1+c_2+12})$ vertices in expectation in the graph G we have constructed; if it is a NO instance, we can infect at most $O(n^{c_1+c_2+11})$ vertices in expectation for any choice of k seeds.*

Proof. If the SETCOVER instance is a YES instance, we are able to choose k subsets $\{A_{i_1}, \dots, A_{i_k}\} \subseteq A$ such that $A_{i_1} \cup \dots \cup A_{i_k} = U$. We choose the k vertices corresponding to these k subsets as seeds.

We say that a clique representing an element is *activated* if all its m vertices are infected with probabilities more than $p^* - \varepsilon$. If a vertex representing a subset is seeded, for each clique representing the element it covers, each of the m vertices in this clique will be infected with probability a_1 . Thus, ma_1 vertices will be infected in expectation. According to Chernoff-Hoeffding inequality, with probability at least $1 - \exp(-\frac{1}{8}a_1^2m)$, there are more than $\frac{1}{2}a_1m$ infected vertices in the clique. If this happens, in the next cascade iteration, each vertex in the clique has more than $\frac{1}{2}a_1m$ infected neighbors, so it will be infected with probability at least $a_{\lfloor \frac{1}{2}a_1m \rfloor} \geq a_{\lfloor a_1n \rfloor} > p^* - \varepsilon$ (notice that $\frac{1}{2}m = \Theta(n^{c_1+c_2+2}) \gg n$). Therefore, if a vertex representing a subset is seeded and a clique representing an element is in this subset, then this clique is activated with probability at least $1 - \exp(-\frac{1}{8}a_1^2n)$.

By our choice of k seeds, each of the clique is activated with probability at least $1 - \exp(-\frac{1}{8}a_1^2n)$. By Boole's inequality, all the n cliques will be activated with probability at least

$$p_{\text{activated}} = 1 - \kappa n \exp\left(-\frac{1}{8}a_1^2n\right) = \Theta(1).$$

In the highly likely case where all the n cliques are activated, all the vertices at the input ends of all the AND gadgets will be infected with probability more than $a_1(p^* - \varepsilon)$. Since the parameter $p_0 = a_1(p^* - \varepsilon)$ is set for the AND gadget, the output vertex v falls into case (2) in Definition 13, which means it will be infected with probability more than $p_2 - \varepsilon_2$. Therefore, all the M_1 vertices connected to v in each of the M_2 copies will be infected with probability at least $a_1(p_2 - \varepsilon_2)$, so the expected total number of infected vertices is at least $p_{\text{activated}} \cdot a_1(p_2 - \varepsilon_2)M_1M_2 = \Theta(n^{c_1+c_2+12})$.

On the other hand, if the SETCOVER instance is a NO instance, consider any choice of k seeds with k_1 of them in the κ vertices representing subsets, k_2 of them in the n cliques, and the remaining $k_3 = k - k_1 - k_2$ of them in the verification part. We first show that at least one clique will not be activated.

The k_3 vertices in the verification part play no role in activating the cliques, as the n cliques are connected to the verification part by directed edges. As for the k_2 vertices in the cliques, since we assume each element in U is in at least one subset, infecting any vertex in any clique is at most as good as infecting the vertex representing the subset covering the element that the clique represents. Therefore, when analyzing the activation of cliques, we can reason as if these k_2 seeds are among the κ subsets. Since the SETCOVER instance is a NO instance and we have picked $k_1 + k_2 \leq k$ subsets, at least one clique will not be activated.

Among the M_2 AND gadgets, at most k_2 of them take the input vertices which are connected from the k_2 seeds in the cliques. Since these k_2 seeds are infected with probability 1 making these input vertices infect with probability a_1 which may be larger than $\frac{11}{10}a_1(p^* - \varepsilon)$, the outputs of these k_2 AND gadgets are unknown as it falls into neither case (1) nor case (2). We have also assumed k_3 seeds are selected in the verification parts, so we also do not know the outputs of another (at most) k_3 AND gadgets.

For the remaining $M_2 - k_2 - k_3$ AND gadgets, they fall into case (1) by the fact that at least one clique is not activated and our setting $p_0 = a_1(p^* - \varepsilon)$ for the AND gadget. Since we have set the AND gadget parameter $\varepsilon_1 = \frac{1}{n}$, the output vertex v will be infected with probability less than $\frac{1}{n}$, which will infect at most $a_1 \frac{M_1}{n}$ vertices in expectation among the M_1 vertices on the right hand side of Figure 4. Notice that each AND gadget has $O(n^{c_1+c_2+1})$ vertices by Lemma 3, and the set cover part has $\kappa + nm$ vertices. In this case, even if all the $\kappa + nm + (M_2 - k_2 - k_3) \cdot O(n^{c_1+c_2+1}) = O(n^{c_1+c_2+3})$ vertices in the set cover part and the $(M_2 - k_2 - k_3)$ AND gadgets are infected, the total number of infected vertices cannot exceed $O(n^{c_1+c_2+3}) + M_2 \cdot a_1 \frac{M_1}{n} = O(n^{c_1+c_2+11})$.

Finally, for those remaining $k_2 + k_3$ AND gadgets whose outputs are unknown, even if all vertices in these $k_2 + k_3$ copies of AND gadgets and their attached M_1 vertices are infected, this total number is still $(k_2 + k_3) \cdot (O(n^{c_1+c_2+1}) + M_1) = (k_2 + k_3) \cdot O(n^{c_1+c_2+10}) = O(n^{c_1+c_2+11})$. Therefore, if the SETCOVER instance is a NO instance, we can infect at most $O(n^{c_1+c_2+11})$ vertices in G . \square

6.2 Proof of Theorem 4 for $a_1 > 0$ with Undirected Graphs

To prove Theorem 4 for undirected graphs, we will need the following *directed edge gadget* which simulates directed edges, and the construction of this gadget also requires the property $a_2 > 2a_1$. This is because the directed edge gadget also uses probability filter gadgets as building blocks.

Definition 15. A $(\Upsilon, \varepsilon, b, f)$ -directed edge gadget $\langle u, v \rangle$ takes one vertex u as input and output one vertex v such that the following properties hold.

1. directed property: If u is connected to each of the Υ vertices v_1, \dots, v_Υ by a directed edge gadget $\langle u, v_i \rangle$, and v_1, \dots, v_Υ are already infected, then u will be infected with probability less than ε .
2. If the input u is infected, then the output v will be infected with probability b . Moreover, $b > 0$.

The size of a directed edge gadget is given by the following lemma.

Lemma 6. For any 2-submodular function f with $a_1 > 0$, any positive integer Υ and any $\varepsilon > 0$, there exists $b \in (0, 1)$ such that we can construct a $(\Upsilon, \varepsilon, b, f)$ -directed edge gadget with $\Theta(\Upsilon^d(1/\varepsilon)^d)$ vertices and $\Theta(\Upsilon^d(1/\varepsilon)^d)$ edges, where $d > 1$ is a constant depending only on f .

We also need the following lemma.

Lemma 7. Given an $(I, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget, for any Υ and ε , we can construct a $(\Upsilon, \varepsilon, b, f)$ -directed edge gadget with $b \in (p_2 - \frac{1}{2}\varepsilon_2, p_2]$ using $\Theta(\Upsilon^d(1/\varepsilon)^d)$ vertices and $\Theta(\Upsilon^d(1/\varepsilon)^d)$ edges, where $d > 1$ is a constant depending only on f .

The construction of the directed edge gadget and the proofs of Lemma 6 and Lemma 7 are deferred to Section 6.6.

6.2.1 A Reduction from SETCOVER

According to Lemma 4, for $I = n$, there exists a constant $p_2 > 0$, such that if we set $\varepsilon_1 = \frac{1}{n}$ and $\varepsilon_2 = \frac{1}{100}p_2$, we can construct an $(n, \Lambda, p^*(p_2 - \varepsilon_2), p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget, where $\Lambda = O((1/\varepsilon_1)^{c_1} n^{c_2}) = O(n^{c_1+c_2})$. Define $M_2 = n^2$ and $m = M_2\Lambda$ as before, and we will define M_1 later.

Applying Lemma 7, we can construct a (mn, m^{-2}, b, f) -directed edge gadget such that $b \in (p_2 - \frac{1}{2}\varepsilon_2, p_2]$. The numbers of vertices and edges in this directed edge gadget are both

$$\Theta\left((mn)^d (1/m^{-2})^d\right) = \Theta\left(m^{3d} n^d\right) = \Theta\left(n^{(3c_1+3c_2+7)d}\right).$$

We will use this directed edge gadget exclusively in the construction. Finally, define $M_1 = n^{(30c_1+30c_2+70)d}$.

We will construct a undirected graph G similar to the one in the last section, with some modifications. We make the following two modifications:

1. We replace all directed edges in Figure 4 by (mn, m^{-2}, b, f) -directed edge gadgets. These consist of 1) the directed edges connecting between the κ vertices representing subsets and the n cliques representing elements and 2) the directed edges connecting between the set cover part and the verification part.
2. We use the $(n, \Lambda, p^*(p_2 - \varepsilon_2), p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadgets in the verification part instead of the $(n, \Lambda, a_1(p^* - \varepsilon), p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadgets.

For the remaining parts of the construction, all the edges, including the ones in the clique, the ones in the AND gadget, and the ones connected to the M_1 vertices on the right hand side of Figure 4, are undirected edges. In particular, we recall that the edges in the AND gadget are undirected.

The Size of the Construction We show that the total number of vertices in G is still of polynomial size. In the set cover part, we have created at most κmn directed edge gadgets between the κ vertices and the mn vertices in the n cliques. The total size of the set cover part is at most $\kappa + mn + \kappa mn \cdot \Theta(n^{(3c_1+3c_2+7)d}) = O(n^{(3c_1+3c_2+7)d+c_1+c_2+4})$.

In the verification part, the AND gadget contains $O(n^{c_1+c_2+1})$ vertices by Lemma 4, the total number of vertices is $M_2(O(n^{c_1+c_2+1}) + M_1) = \Theta(n^{(30c_1+30c_2+70)d+2})$.

Since $d > 1$ by Lemma 4, N is dominated by the number of vertices in the verification part: $N = \Theta(n^{(30c_1+30c_2+70)d+2})$.

Finally, Theorem 4 for undirected graphs follows immediately from the following lemma.

Lemma 8. *If the SETCOVER instance is a YES instance, by choosing k seeds appropriately, we can infect $\Theta(n^{(30c_1+30c_2+70)d+2})$ vertices in expectation in the graph G ; if it is a NO instance, we can infect at most $O(n^{(30c_1+30c_2+70)d+1})$ vertices in expectation for any choice of k seeds.*

Proof. If the SETCOVER instance is a YES instance, we are able to choose k subsets $\{A_{i_1}, \dots, A_{i_k}\} \subseteq A$ such that $A_{i_1} \cup \dots \cup A_{i_k} = U$. We choose the k vertices corresponding to these k subsets as the seeds. Since the SETCOVER instance is a YES instance, for each clique, each vertex is connected from a seed by a directed edge gadget, which will be infected with probability b . In each clique, bm vertices will be infected in expectation, and the remaining vertices in the clique will be infected with probability at least $a_{\lfloor bm \rfloor}$, which has limit p^* as $n \rightarrow \infty$.

By the same analysis in the proof of Lemma 5, with a high probability $p_{\text{activated}} = \Theta(1)$, all the n cliques will be activated such that all vertices in the clique will be infected with probability $p^* - \varepsilon$ for certain $\varepsilon = o(1)$. By our construction and Lemma 7, each of the mn vertices that are passed into the input of the AND gadget will be infected with probability

$$(p^* - \varepsilon)b \in \left((p^* - \varepsilon) \left(p_2 - \frac{1}{2}\varepsilon_2 \right), (p^* - \varepsilon)p_2 \right] \subseteq \left(p^*(p_2 - \varepsilon_2), \frac{11}{10}p^*(p_2 - \varepsilon_2) \right),$$

by noticing that $\varepsilon = o(1)$ and $\varepsilon_2 = \frac{1}{100}p_2 < \frac{1}{10}p_2$ is a constant. Thus, the AND gadget falls into case (2) of Definition 13, so the output vertex v of the AND gadget will be infected with probability more than $p_2 - \varepsilon_2$. Therefore, each of the M_1 vertices will be infected with probability

$p_{\text{activated}} a_1(p_2 - \varepsilon_2)$, and the expected total number of infected vertices in those M_2 copies of M_1 vertices is already $p_{\text{activated}} a_1(p_2 - \varepsilon_2) M_1 M_2 = \Theta(n^{(30c_1+30c_2+70)d+2})$.

If the SETCOVER instance is a NO instance, consider any choice of the k seeds with k_1 seeds in the κ vertices representing subsets, k_2 seeds in the directed edge gadgets connecting the κ vertices and nm vertices in the n cliques, k_3 seeds in the n cliques, k_4 seeds in the directed edge gadgets between the n cliques in the set cover part and the inputs of the AND gadget in the verification part, and the remaining $k_5 = k - k_1 - k_2 - k_3 - k_4$ seeds in the verification parts. We first aim to show that at least one clique will not be activated with high probability.

When analyzing cliques activation, it is easy to see that putting k_2 seeds on the directed edge gadgets is at most as good as putting them on the corresponding vertices representing the subsets. Similarly, putting k_4 seeds on the directed edge gadgets connecting the set cover part and the verification part is at most as good as putting them on the corresponding vertices in the cliques, and having k_3+k_4 seeds in the cliques is at most as good as having them in the κ vertices representing the subsets covering the elements that those cliques represent. Thus, we can reason as if we have selected $k_1 + k_2 + k_3 + k_4$ subsets in the SETCOVER problem. Since the SETCOVER instance is a NO instance, those $k_1 + k_2 + k_3 + k_4 \leq k$ seeds cannot cover all the cliques. As for the k_5 seeds in the verification part, their influences on each vertex in the n cliques is at most m^{-2} based on Definition 15, which has remote effect to the cliques, and we will discuss about it later.

To show that at least one clique is not activated, it remains to show that the clique not covered by those $k_1 + k_2 + k_3 + k_4$ vertices cannot be activated. For each of those vertices representing subsets that are not picked, since it is connected to at most mn vertices (m vertices in each of the n cliques) by the (mn, m^{-2}, b, f) -directed edge gadgets, it will be infected with probability at most m^{-2} by Definition 15. For each vertex in each uncovered clique, it may only be infected due to 1) the influence from one of the κ vertices which is not seeded and which is infected with probability at most m^{-2} , or 2) the influence from the k_5 seeds from the verification parts. In particular, it will be infected due to (1) with probability bm^{-2} , and it will be infected due to (2) with probability m^{-2} . By Boole's inequality, the probability that there exist infected vertices in a uncovered clique is at most $m \cdot (bm^{-2} + m^{-2}) = O(m^{-1})$. Since there can be at most n uncovered cliques, the probability that all uncovered cliques contain no infected vertex is at least

$$p_{\text{no}} = 1 - n \cdot O(m^{-1}) > 1 - O\left(\frac{1}{n}\right).$$

Therefore, with the probability above, there exists at least one clique which is not activated.

In the case that not all cliques are activated, since all the vertices in a not activated clique are infected with probability 0, the corresponding input vertices to the AND gadget are also infected with probability $0b = 0$. The output vertices v in the AND gadgets therefore fall into case (1) in at least $M_2 - k_3 - k_4 - k_5$ copies. Thus, in each of the corresponding $M_2 - k_3 - k_4 - k_5$ copies of the M_1 vertices bundle (on the rightmost of Figure 4), the expected number of infected vertices is at most $\varepsilon_1 \cdot M_1 = O(n^{(30c_1+30c_2+70)d-1})$. In this case, even if all the vertices in the entire set cover part, the mn directed edge gadgets connecting the two parts, all the M_2 AND-gadgets, and the remaining $k_3 + k_4 + k_5$ copies of the M_1 vertices bundles, the total number of infected vertices

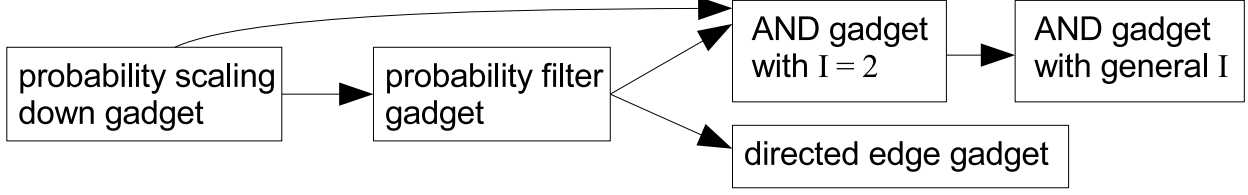


Figure 5: The relation of all the gadgets defined

is at most

$$\begin{aligned}
& \kappa mn \cdot \Theta \left(n^{(3c_1+3c_2+7)d} \right) + mn \cdot \Theta \left(n^{(3c_1+3c_2+7)d} \right) + M_2 \cdot O \left(n^{c_1+c_2+1} \right) \\
& + (k_3 + k_4 + k_5) \left(O \left(n^{c_1+c_2+1} \right) + M_1 \right) + (M_2 - k_3 - k_4 - k_5) O \left(n^{(30c_1+30c_2+70)d-1} \right) \\
= & O \left(n^{(3c_1+3c_2+7)d+c_1+c_2+4} \right) + \Theta \left(n^{(3c_1+3c_2+7)d+3} \right) + O \left(n^{c_1+c_2+3} \right) \\
& + O \left(n^{(30c_1+30c_2+70)d+1} \right) + O \left(n^{(30c_1+30c_2+70)d+1} \right) \\
= & O \left(n^{(30c_1+30c_2+70)d+1} \right).
\end{aligned}$$

Finally, even assuming all vertices in G are infected in the case that all cliques are activated (which happens with probability $1 - p_{\text{no}} < O\left(\frac{1}{n}\right)$), the expected number of infected vertices is at most

$$p_{\text{no}} \cdot O \left(n^{(30c_1+30c_2+70)d+1} \right) + (1 - p_{\text{no}})N = O \left(n^{(30c_1+30c_2+70)d+1} \right),$$

which concludes the lemma. \square

6.3 Constructions of Some Other Required Gadgets

Before constructing the AND gadget and the directed edge gadget, we need some other gadgets. In this section and the next two sections, graph with undirected edges are considered.

We will construct the *probability scaling down gadget* and the *probability filter gadget*, which are used to construct the AND gadget and the directed edge gadget. The relation of these gadgets are shown in Figure 5.

6.3.1 Probability Scaling Down Gadget

We first define and construct the following *probability scaling down gadget* which is an essential component of both the AND gadget and the directed edge gadget.

Definition 16. *The (α, ε, f) -probability scaling down gadget takes one vertex u as input and output a vertex v such that*

- *if u is infected with probability p_u , v will be infected with probability $p_v \in (\alpha p_u - \varepsilon, \alpha p_u]$.*

Lemma 9. *For any 2-submodular function f with $a_1 > 0$, any constant $\varepsilon > 0$ and any α with $0 < \alpha \leq p^*$, there exists an (α, ε, f) -probability scaling down gadget with constant numbers of vertices and edges.*

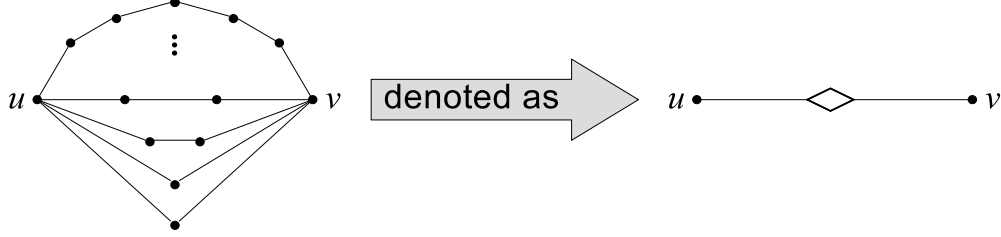


Figure 6: The probability scaling down gadget

Proof. To construct this gadget, we iteratively add paths from u to v , where a path of length ℓ consists of $\ell - 1$ vertices $w_1, \dots, w_{\ell-1}$ and ℓ edges $(u, w_1), (w_1, w_2), \dots, (w_{\ell-1}, v)$. Given p_u , by repeatedly adding paths from u to v , we are increasing p_v . In each iteration i , we add a path of length ℓ_i from u to v , where ℓ_i is the minimum length to maintain $p_v \leq \alpha p_u$. That is, either it is true that $p_v > \alpha p_u$ if a path of length $\ell_i - 1$ was added, or $\ell_i = 2$ which is already the minimum length a path can ever be. The iterative process ends if $p_v \in (\alpha p_u - \varepsilon, \alpha p_u]$, and it is straightforward to check that such process will end as long as $\alpha \in (0, p^*]$. Figure 6 illustrates the probability scaling down gadget.

The size of the probability scaling down gadget depends on the influence function f and the small constant ε . Since f is fixed in advance, the size of this gadget is constant. \square

Remark 1. *The probability scaling down gadget is symmetric. Given $p_v = \alpha p_u$, then $p_u = \alpha p_v$ if v becomes the input and u becomes the output.*

6.3.2 Probability Filter Gadget

Based on the probability scaling down gadget, we can construct the following *probability filter gadget*.

Definition 17. *A $(\Lambda, p_1, p_2, \varepsilon_1, \varepsilon_2, f)$ -probability filter gadget takes Λ vertices as input, and outputs a vertex such that*

1. *if each vertex in the Λ inputs is infected independently with a same probability less than p_1 , then the vertex on the output end will be infected with a probability less than ε_1 ;*
2. *if each vertex in the Λ inputs is infected independently with a same probability in $(p_1, p_2]$, then the vertex on the output end will be infected with a probability in $(p_2 - \varepsilon_2, p_2]$.*

We aim to show the following lemma in this subsection.

Lemma 10. *Given any 2-submodular influence function f with $a_1 > 0$, any constant $\varepsilon_2 > 0$, any $\varepsilon_1 > 0$, and any ratio $r > 0$, we can construct a $(\Lambda, p_1, p_2, \varepsilon_1, \varepsilon_2, f)$ -probability filter gadget with $p_2/p_1 > r$ and $\Lambda = O((1/\varepsilon_1)^c)$, and this probability filter gadget contains $O((1/\varepsilon_1)^c)$ vertices and $O((1/\varepsilon_1)^c)$ edges, where c is a constant.*

To construct the probability filter gadget, we first construct the gadget shown in Figure 7, which is the building block of this gadget. We will call this building block *probability separation block*. As shown in the figure, this building block takes h vertices as input and outputs one vertex. Particularly, we apply h probability scaling down gadgets to “scale down” the probabilities of all input vertices’ infection by a factor of α , and then connect those vertices to the output vertex.

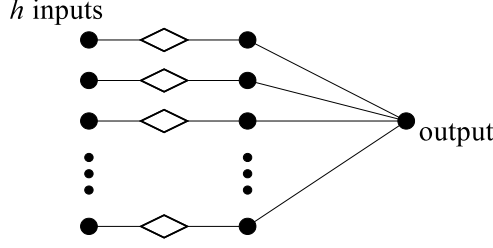


Figure 7: The probability separation block

The probability filter gadget consists of ℓ layers such that the i -th layer consists of $h^{\ell-i}$ such probability separation blocks, where the output vertices of every h probability separation blocks in the i -th layer are the input of a probability separation block in the $(i+1)$ -th layer. Because there are $h^{\ell-1}$ probability separation blocks in the first layer, the probability filter gadget takes $\Lambda = h^\ell$ vertices as input. The probability filter gadget outputs a single vertex after ℓ layers. We will tune the value of α , h and ℓ such that the two properties in Definition 17 hold for certain thresholds p_1 and p_2 .

For each probability separation block, suppose each of the h vertices in the input are infected with probability x independently, and let $y = y(x)$ be the probability that the output vertex is infected. We aim to tune the value of α and h such that the graph of $y(x)$ looks like Figure 8.

By considering the number of infected neighbors of the output vertex, it is straightforward to see that

$$y = \sum_{i=1}^h \binom{h}{i} a_i (\alpha x)^i (1 - \alpha x)^{h-i}. \quad (4)$$

For sufficiently small x , we have

$$y = h\alpha a_1 x + \frac{h(h-1)}{2} \alpha^2 (a_2 - 2a_1) x^2 + o(x^2).$$

Choosing a sufficiently small constant $\delta > 0$ and choosing α (h will be set in the future) to satisfy $h\alpha a_1 = 1 - \delta$, we have

$$y - x = -\delta x + \frac{h(h-1)}{2} \alpha^2 (a_2 - 2a_1) x^2 + o(x^2).$$

Since $y - x = -\delta x + o(x)$, we can see that $y < x$ for small enough x . On the other hand, for sufficiently large h and sufficiently small δ (and adjusting α such that $h\alpha a_1 = 1 - \delta$ still holds¹²), we have

$$\frac{h(h-1)}{2} \alpha^2 = \frac{1}{2} h^2 \alpha^2 - \frac{h}{2} \alpha^2 = \frac{(1-\delta)^2}{2a_1^2} - \frac{(1-\delta)^2}{2ha_1^2} > \frac{1}{3a_1^2}.$$

¹²According to Definition 16 and Lemma 9, given the scale α^* for which we want to adjust to, we can construct a probability scaling down gadget such that the actual scale α is arbitrarily close to α^* . Although we cannot make the adjustment exact, a close enough approximation would still satisfy our purpose here, as all we want is δ to be small enough, or $h\alpha a_1$ to be close enough to 1.

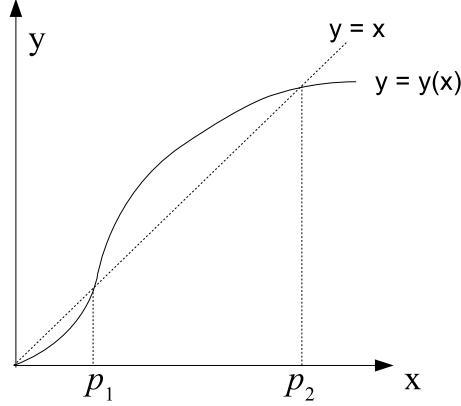


Figure 8: The output probability y versus the input probability x

We can see from the following that $y > x$ after a while as x increases.

$$\begin{aligned}
 x_1 = \frac{6a_1^2}{a_2 - 2a_1}\delta &\implies y(x_1) - x_1 > -\delta x_1 + \frac{a_2 - 2a_1}{3a_1^2}x_1^2 + o(x_1^2) \\
 &= \frac{6a_1^2}{a_2 - 2a_1}\delta^2 + o(\delta^2) \\
 &> 0.
 \end{aligned}$$

Notice that the 2-submodularity of f makes sure $a_2 > 2a_1$ such that x_1 is positive.

We have seen that $y < x$ for small enough x , and $y > x$ after x increases. There must be a threshold p_1 such that $y = x$ at $x = p_1$ by the Intermediate Value Theorem. On the other hand, y is upper bounded by p^* while x can be as large as 1, so $y \leq x$ for sufficiently large x . The Intermediate Value Theorem suggests there exists another threshold $x = p_2 > p_1$ such that $y = x$. Consequently, Figure 8 indeed represents the graph of $y(x)$ for the proper choices of α and h .

Finally, from the graph in Figure 8, we can see that the infection probability of the output vertices in the i -th layer increases as i increases, if all the $\Lambda = h^\ell$ input vertices are infected with an independent probability larger than p_1 . In contrast, the infection probability of the output vertices in the i -th layer decreases as i increase, if all the $\Lambda = h^\ell$ input vertices are infected with an independent probability less than p_1 . By setting ℓ large enough, we can make both (1) and (2) in Definition 17 hold.

Before we move on, we show some properties of the thresholds p_1 and p_2 , and our objective is to show the following proposition which is a part of Lemma 10.

Proposition 3. *For any large ratio $r > 0$, we can find h and α such that $p_2/p_1 > r$.*

By the calculation above, the proposition below follows immediately.

Proposition 4. $p_1 < \frac{6a_1^2}{a_2 - 2a_1}\delta$.

We also have the following lower bound for p_2 .

Proposition 5. *By choosing h sufficiently large and δ sufficiently small, we have $p_2 > a_1\gamma$ for any γ such that*

$$a_2(1 - e^{-\gamma} - \gamma e^{-\gamma}) - a_1(\gamma - \gamma e^{-\gamma}) > 0.$$

Proof. By replacing all a_3, a_4, \dots, a_h to a_2 in Equation (4), we have

$$\begin{aligned}
y &\geq \sum_{i=1}^h \binom{h}{i} a_2 (\alpha x)^i (1 - \alpha x)^{h-i} - \binom{h}{1} (a_2 - a_1) \alpha x (1 - \alpha x)^{h-1} \\
&= a_2 \left(\sum_{i=0}^h \binom{h}{i} (\alpha x)^i (1 - \alpha x)^{h-i} - (1 - \alpha x)^h \right) - h(a_2 - a_1) \alpha x (1 - \alpha x)^{h-1} \\
&= a_2 - a_2 (1 - \alpha x)^h - h(a_2 - a_1) \alpha x (1 - \alpha x)^{h-1} \\
&= a_2 - a_2 \exp(h \ln(1 - \alpha x)) - h(a_2 - a_1) \alpha x \exp((h-1) \ln(1 - \alpha x)) \\
&\geq a_2 - a_2 \exp(-h\alpha x) - h(a_2 - a_1) \alpha x \exp(-\alpha x(h-1)). \quad (\text{concavity of } \ln \text{ function})
\end{aligned}$$

Letting $x = a_1 \gamma$, we have

$$\begin{aligned}
y - x &\geq a_2 - a_2 \exp(-\gamma(1 - \delta)) - (a_2 - a_1)(1 - \delta) \gamma \exp\left(\gamma(1 - \delta) \left(\frac{1}{h} - 1\right)\right) - a_1 \gamma \\
&\quad (\text{since } x = a_1 \gamma \text{ and } h\alpha a_1 = 1 - \delta) \\
&> a_2(1 - e^{-\gamma} - \gamma e^{-\gamma}) - a_1(\gamma - \gamma e^{-\gamma}) - \epsilon,
\end{aligned}$$

where in the last step, for any $\epsilon > 0$, we can find small enough δ and large enough h to make the inequality holds. Rigorously, we have $1 - \delta \rightarrow 1$ and $\frac{1}{h} \rightarrow 0$ for $\delta \rightarrow 0$ and $h \rightarrow \infty$. The expression in the second last step is a continuous function, which has limit $a_2(1 - e^{-\gamma} - \gamma e^{-\gamma}) - a_1(\gamma - \gamma e^{-\gamma})$, and the last step is obtained by the definition of limit.

Therefore, $y - x > 0$ for any $x > p_1$ with $x = a_1 \gamma$, where γ satisfies

$$a_2(1 - e^{-\gamma} - \gamma e^{-\gamma}) - a_1(\gamma - \gamma e^{-\gamma}) > 0,$$

which implies the proposition. \square

We remark that there always exists γ satisfying the inequality in Proposition 5. To see this, we show that $\Phi(\gamma) := a_2(1 - e^{-\gamma} - \gamma e^{-\gamma}) - a_1(\gamma - \gamma e^{-\gamma}) > 0$ when γ is sufficiently small. By straightforward calculations, we have $\Phi(0) = \Phi'(0) = 0$ and $\Phi''(0) = a_2 - 2a_1 > 0$, which means $\Phi(0) = 0$ and Φ is increasing on $[0, \gamma_0)$ for some small γ_0 , which further implies that Φ is positive on $[0, \gamma_0)$.

Proposition 4 implies that we can construct the probability filter gadget with arbitrarily small p_1 by setting δ small. On the other hand, Proposition 5 implies that p_2 can be made larger than some number depending only on a_1 and a_2 , which in particular can be considerably larger than p_1 , which yields Proposition 3.

Finally, we are ready to show Lemma 10.

Proof of Lemma 10. The possibility of this construction is straightforward, as the construction is already made explicit in this section. It remains to show that the gadget contains $O((1/\varepsilon_1)^c)$ vertices and $O((1/\varepsilon_1)^c)$ edges, and $\Lambda = O((1/\varepsilon_1)^c)$.

Since ε_2 is a constant, we only need constantly many layers such that the input probability x increases to more than $p_2 - \varepsilon_2$, if x is initially larger than p_1 .

To investigate how many layers are needed to make x decreases to less than ε_1 in the case x is initially smaller than p_1 , recall that in each layer of the probability filter gadget, the input probability x is updated to y such that $y - x = -\delta x + o(x)$ for sufficiently small x , so each time x

is decreased by a factor of $(1 - \delta)$. After a constant number of layers, x will be sufficiently small such that the term $o(x)$ is negligible, and after another $\frac{\log(1/\varepsilon_1)}{\log(1/(1-\delta))}$ layers, x will decrease by a factor of $(1 - \delta)^{\frac{\log(1/\varepsilon_1)}{\log(1/(1-\delta))}} = \varepsilon_1$, which makes the value of x much smaller than ε_1 . Therefore, we need at most $\ell = O(\log(1/\varepsilon_1))$ layers. Let χ_v, χ_e be the number of vertices and edges respectively in a probability separation block shown in Figure 7, and they are both constants according to Lemma 9. The total number of vertices in a probability filter gadget is

$$\sum_{i=1}^{\ell} \chi_v \cdot h^{\ell-i} = \chi_v \frac{h^{\ell} - 1}{h - 1} = \Theta(h^{\ell}) = O((1/\varepsilon_1)^c),$$

and the total number of edges has the same asymptotic bound by the same calculation above, with χ_v changed to χ_e . Thus, we conclude that the gadget contains $O((1/\varepsilon_1)^c)$ vertices and $O((1/\varepsilon_1)^c)$ edges.

For Λ , we have $\Lambda = h^{\ell} = O((1/\varepsilon_1)^c)$ by our construction, which concludes the last part of the lemma. \square

6.4 Construction of the AND Gadget with $I = 2$

In this section, we construct the AND gadget with parameter $I = 2$. The AND gadget makes use of a single probability filter gadget with the same choices of parameters $\Lambda, p_2, \varepsilon_1, \varepsilon_2$ and f . The AND gadget takes two sets I_1, I_2 of vertices as inputs, and each set has $\Lambda = h^{\ell}$ vertices. Let $I_1 = \{u_1, u_2, \dots, u_{\Lambda}\}$ and $I_2 = \{v_1, v_2, \dots, v_{\Lambda}\}$. We create Λ vertices $w_1, w_2, \dots, w_{\Lambda}$ and create two edges $(u_i, w_i), (v_i, w_i)$ for each $i = 1, 2, \dots, \Lambda$. We apply the probability scaling down gadgets to create another Λ vertices $w'_1, w'_2, \dots, w'_{\Lambda}$ such that $p(w'_i) = \beta p(w_i)$ for each $i = 1, 2, \dots, \Lambda$, where β is set to the value such that

$$\beta \varphi_T^+(p_0) < p_2, \quad \beta \varphi_T^-(p_0) > p_1, \quad \text{and} \quad \beta \varphi_F^+(p_0) < p_1,$$

where

$$\begin{aligned} \varphi_T^+(p_0) &= (a_2 - 2a_1) \left(\frac{11}{10} p_0 \right)^2 + 2a_1 \left(\frac{11}{10} p_0 \right), \\ \varphi_T^-(p_0) &= (a_2 - 2a_1) p_0^2 + 2a_1 p_0, \\ \varphi_F^+(p_0) &= \frac{11}{20} (a_2 - 2a_1) p_0^2 + \frac{16}{10} a_1 p_0. \end{aligned}$$

The construction is shown in Figure 9.

Notice that if all u_i and v_i are infected with an independent probability in the interval $(p_0, \frac{11}{10} p_0)$, that is, the inputs I_1, I_2 fall into case (2) in Definition 13, w_i will be infected with probability

$$\begin{aligned} p(w_i) &= a_2 p(u_i) p(v_i) + a_1 p(u_i) (1 - p(v_i)) + a_1 p(v_i) (1 - p(u_i)) \\ &= (a_2 - 2a_1) p(u_i) p(v_i) + a_1 p(u_i) + a_1 p(v_i), \end{aligned}$$

which is in the interval $(\varphi_T^-(p_0), \varphi_T^+(p_0))$.

On the other hand, if one of u_i and v_i is infected with probability less than $\frac{1}{2} p_0$ and the other one is infected with probability less than $\frac{11}{10} p_0$, that is, the inputs I_1, I_2 fall into case (1) in Definition 13,

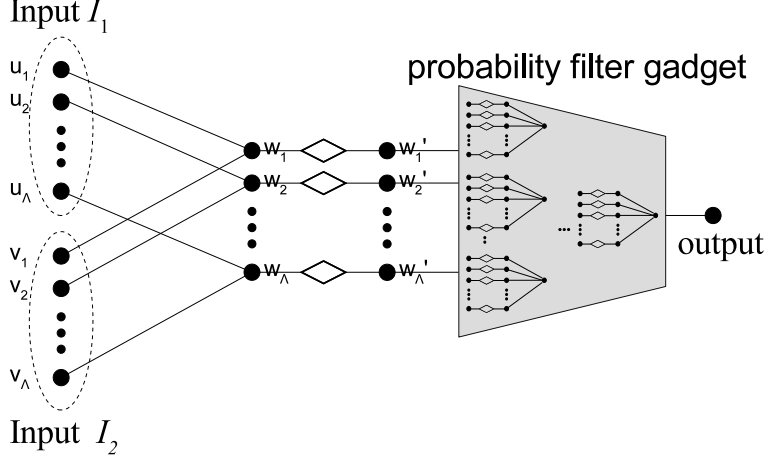


Figure 9: AND gadget with $I = 2$

w_i will be infected with probability

$$\begin{aligned}
 p(w_i) &= (a_2 - 2a_1)p(u_i)p(v_i) + a_1p(u_i) + a_1p(v_i) \\
 &< \frac{11}{20}(a_2 - 2a_1)p_0^2 + \frac{16}{10}a_1p_0 \\
 &= \varphi_F^+(p_0).
 \end{aligned}$$

Given that $(\beta\varphi_T^-(p_0), \beta\varphi_T^+(p_0)) \subseteq (p_1, p_2)$ and $\beta\varphi_F^+(p_0) < p_1$, it is now straightforward to check that the two properties (1) and (2) in Definition 13 hold for $I = 2$, since the probability filter gadget will “filter” the two probabilities such that one goes to a value less than ε_1 and the other goes into $(p_2 - \varepsilon_2, p_2]$.

By our construction of probability scaling down gadget, the factor must satisfy $\beta \leq p^*$. It seems worrying that $(\varphi_T^-(p_0), \varphi_T^+(p_0))$ and $\varphi_F^+(p_0)$ will be both scaled down to smaller than p_1 even if we take maximum $\beta = p^*$. Indeed, Proposition 3 and Proposition 4 ensure that this cannot happen, as we can always make p_1 small enough by making δ small enough. We remark here that the choice of δ depends on p_0 and p^* (it needs to be considerably smaller than some polynomial of p_0 such that $(\varphi_T^-(p_0), \varphi_T^+(p_0))$ and $\varphi_F^+(p_0)$ can be scaled down to different sides of p_1), where p^* depends only on f .

Now we prove the following lemma, which is a special case of Lemma 3 with $I = 2$.

Lemma 11. *Given any 2-submodular function f with $a_1 > 0$ and any constant threshold $p_0 > 0$, there exists a constant $p_2 > 0$ depending on p_0 and f such that for any constant $\varepsilon_2 > 0$ and any $\varepsilon_1 > 0$, we can construct a $(2, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget with $\Lambda = O((1/\varepsilon_1)^{c_1})$, and the number of vertices and edges in this AND gadget are both $O((1/\varepsilon_1)^{c_1})$, where c_1 is a constant.*

Proof. The existence of this AND gadget is shown by the explicit construction in this section.

To show that p_2 only depends on p_0 and f , notice that it depends on h, δ and f (in particular, a_1 and a_2 only) according to Proposition 5. Additionally, h, α are selected such that $\delta = 1 - h\alpha a_1$ is small enough, and we have remarked just now that δ depends on p_0 and f . Therefore, p_2 only depends on p_0 and f , as the graph $y = y(x)$ determines the value of p_2 .

For the size of this AND gadget and the input size Λ , the size of this AND gadget is the size of a probability filter gadget plus 3Λ for those u_i, v_i, w_i , and the size of each of both input sets is Λ . Therefore, Lemma 10 implies the second part of this lemma. \square

Remark 2 (Remark of Lemma 11). *Lemma 11 shows that when constructing a $(2, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget, we are free to setup the parameter p_0 , and the parameter p_2 will be determined. After p_2 is determined, we are still free to choose $\varepsilon_1, \varepsilon_2$, and Λ will be then determined. In fact, the two parameters $\varepsilon_1, \varepsilon_2$ decides the number of layers needed in the probability filter gadget, and we can achieve (1) and (2) in Definition 13 for any valid function $y(x)$ with two intersections to the line $y = x$ as it is in Figure 8. That is the reason why we can choose $\varepsilon_1, \varepsilon_2$ after p_2 is determined. In particular, for the same function $y(x)$ but different $\varepsilon_1, \varepsilon_2$, we just need the AND gadgets with different numbers of layers in their inner probability filter gadgets. We will make use of this observation to construct AND gadgets with the same parameters p_0, p_2, f but different $\varepsilon_1, \varepsilon_2$ in the next section.*

To conclude this section, we show that we can also construct a $(2, \Lambda, p_2 - \varepsilon_2, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget and a $(2, \Lambda, p^*(p_2 - \varepsilon_2), p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget which will be used in the next section. Notice that Lemma 11 does not imply the possibility of constructing this AND gadget, as p_2 's existence is supposed to depend on the third parameter, which now become $p_2 - \varepsilon_2$ and $p^*(p_2 - \varepsilon_2)$, two constants related to p_2 .

Lemma 12. *Given any 2-submodular influence function f and any constant threshold $p_0 > 0$, we can construct a $(2, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget, a $(2, \Lambda, p_2 - \varepsilon_2, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget and a $(2, \Lambda, p^*(p_2 - \varepsilon_2), p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget with the same parameters $\Lambda, p_2, \varepsilon_1, \varepsilon_2$.*

Proof. The three AND gadgets are only different at the third parameter, which is the input threshold determining which of the two cases (1) and (2) in Definition 13 the inputs fall into. By our construction, we can use the same structure for the three AND gadgets, except that we use three different scaling down factors $\beta_1, \beta_2, \beta_3$ for the different thresholds $p_0, p_2 - \varepsilon_2$ and $p^*(p_2 - \varepsilon_2)$. In particular, the three probability filter gadgets inside the three AND gadgets can be exactly the same, provided that the ‘‘gap’’ p_2/p_1 is large enough such that

- $(\beta_1\varphi_T^-(p_0), \beta_1\varphi_T^+(p_0))$ and $\beta_1\varphi_F^+(p_0)$ are on the different sides of p_1 ,
- $(\beta_2\varphi_T^-(p_2 - \varepsilon_2), \beta_2\varphi_T^+(p_2 - \varepsilon_2))$ and $\beta_2\varphi_F^+(p_2 - \varepsilon_2)$ are on the different sides of p_1 , and
- $(\beta_3\varphi_T^-(p^*(p_2 - \varepsilon_2)), \beta_3\varphi_T^+(p^*(p_2 - \varepsilon_2)))$ and $\beta_3\varphi_F^+(p^*(p_2 - \varepsilon_2))$ are on the different sides of p_1 .

We know that this is always possible by Proposition 3.

As the same probability filter gadget is used in the two AND gadgets, the four parameters $\Lambda, p_2, \varepsilon_1, \varepsilon_2$, which are inherited from the probability filter gadget by our construction, are identical for the three AND gadgets. \square

6.5 Construction of the AND Gadget with General I of an Integer Power of 2

In this section, we construct the AND gadget in Definition 13 with general I that is an integer power of 2.

A $(I, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget is a $(\log_2 I)$ -level AND circuit using 2-set-input AND gadgets constructed in the previous section as building block. We will use three different types of 2-set-input AND gadgets.

- Type A: $(2, \Lambda_0, p_0, p_2, \frac{1}{3}(p_2 - \varepsilon_2), \varepsilon_2, f)$ -AND gadget.
- Type B: $(2, \Lambda_0, p_2 - \varepsilon_2, p_2, \frac{1}{3}(p_2 - \varepsilon_2), \varepsilon_2, f)$ -AND gadget.
- Type C: $(2, \Lambda_C, p_2 - \varepsilon_2, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget.

Lemma 12 indicates that we can construct A and B , and by Lemma 11 Λ_0 is a constant since $\frac{1}{3}(p_2 - \varepsilon_2)$ is a constant. By Lemma 11 and its remark, we can construct C based on B by adjusting the number of layers in the inner probability filter gadget, and $\Lambda_C = O((1/\varepsilon_1)^{c_1})$ for some constant c_1 .

Figure 10 shows the construction of this AND gadget. The type and the number of AND gadgets in each of the $\log_2 I$ levels are set as follow:

- Level ($\log_2 I$): A single AND gadget of Type C is constructed.
- Level ($\log_2 I - 1$): 2 groups of Λ_C Type B AND gadgets are constructed, and the output vertices in each group are connected to each of the input ends I_1, I_2 of the AND gadget in Level ($\log_2 I$).
- Level ($\log_2 I - 2$): 2^2 groups of $\Lambda_0 \Lambda_C$ Type B AND gadgets are constructed, and the output vertices in each group are connected to each of the input ends I_1, I_2 of the AND gadgets in each of the 2 groups in Level ($\log_2 I - 1$).
- Level ($\log_2 I - 3$): 2^3 groups of $\Lambda_0^2 \Lambda_C$ Type B AND gadgets are constructed, and the output vertices in each group are connected to each of the input ends I_1, I_2 of the AND gadgets in each of the 2^2 groups in Level ($\log_2 I - 2$).
- ...
- Level 2: $2^{\log_2 I - 2}$ groups of $\Lambda_0^{\log_2 I - 3} \Lambda_C$ Type B AND gadgets are constructed, and the output vertices in each group are connected to each of the input ends I_1, I_2 of the AND gadgets in each of the $2^{\log_2 I - 3}$ groups in Level 3.
- Level 1: $2^{\log_2 I - 1}$ groups of $\Lambda_0^{\log_2 I - 2} \Lambda_C$ Type A AND gadgets are constructed, and the output vertices in each group are connected to each of the input ends I_1, I_2 of the AND gadgets in each of the $2^{\log_2 I - 2}$ groups in Level 2.

Finally, the two input sets I_1, I_2 in each of the $2^{\log_2 I - 1} = \frac{I}{2}$ AND gadget groups in Level 1 form two of the I input sets for the $(I, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget we are constructing, and the output vertex of the Type C AND gadget in Level ($\log_2 I$) is the output of the $(I, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget.

We now show that (1) and (2) in Definition 13 hold.

1. If all the vertices in all I input sets are infected with independent probabilities less than $\frac{11}{10}p_0$, and the infection probabilities of the vertices in at least one set are less than $\frac{1}{2}p_0$, then the Type A AND gadgets in at least one group in Level 1 will output vertices with infection probabilities less than $\frac{1}{3}(p_2 - \varepsilon_2)$. Since the threshold (the third parameter) of Type B AND gadgets is set to $(p_2 - \varepsilon_2)$ and $\frac{1}{3}(p_2 - \varepsilon_2) < \frac{1}{2}(p_2 - \varepsilon_2)$, the Type B AND gadgets in at least one group in each of Level 2, 3, ..., $\log_2 I - 1$ will output vertices with infection probabilities less than $\frac{1}{3}(p_2 - \varepsilon_2)$. Finally, at least one of the two input sets for the Type C AND gadget in Level ($\log_2 I$) will be infected with probabilities less than $\frac{1}{3}(p_2 - \varepsilon_2)$, which is less than $\frac{1}{2}(p_2 - \varepsilon_2)$. Thus, the output of the entire $(I, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget is a vertex with infection probabilities less than ε_1 , which implies (1) in Definition 13.
2. If all the vertices in all I input sets are infected with independent probabilities in $(p_0, \frac{11}{10}p_0)$, all the Type A AND gadgets in Level 1 will output vertices with infection probabilities in

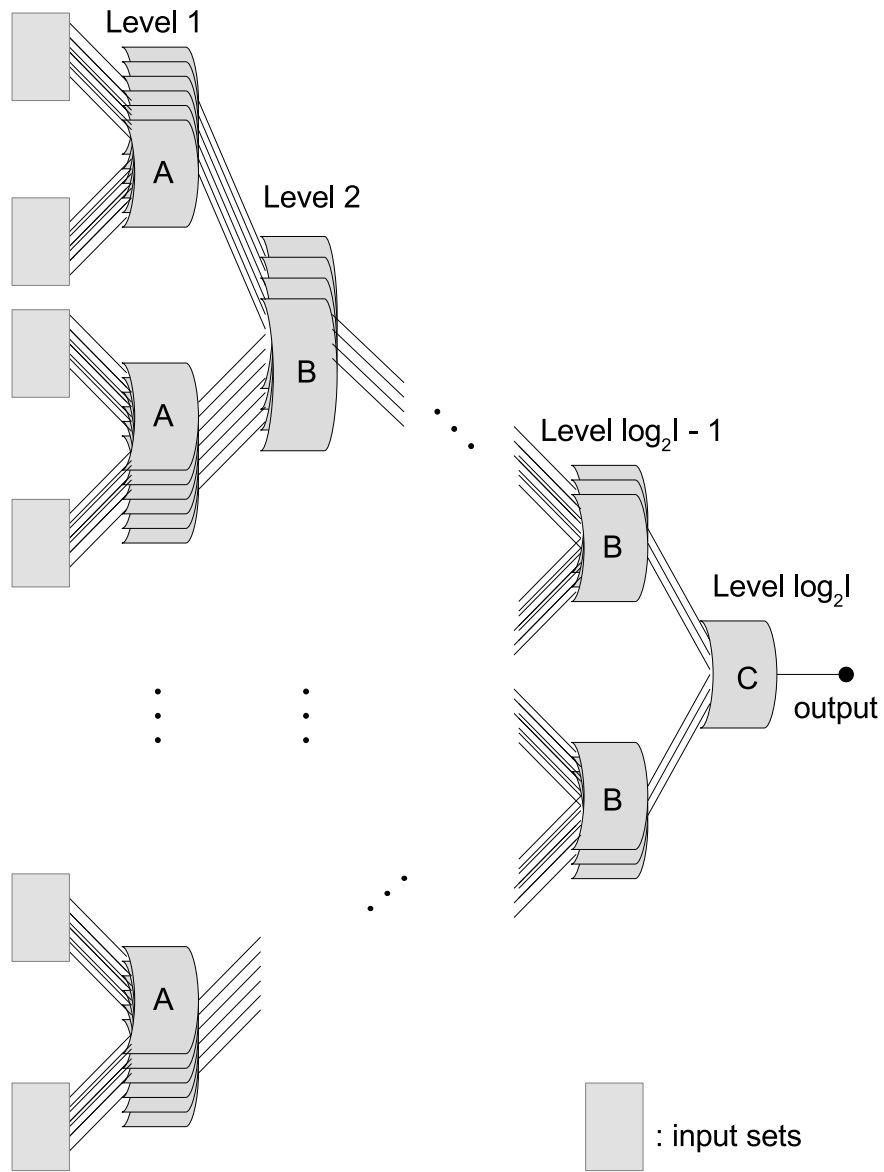


Figure 10: The $(I, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget

$(p_2 - \varepsilon_2, p_2]$. Since $(p_2 - \varepsilon_2, p_2] \subseteq (p_2 - \varepsilon_2, \frac{11}{10}(p_2 - \varepsilon_2))$ for small enough ε_2 ,¹³ all the Type *B* AND gadgets in each of Level 2, 3, \dots , $\log_2 I - 1$ will output vertices with infection probabilities in $(p_2 - \varepsilon_2, p_2]$. Finally, the Type *C* AND gadget in Level $(\log_2 I)$ will output a vertex with infection probability in $(p_2 - \varepsilon_2, p_2]$.

Finally, we prove Lemma 3 and Lemma 4 in Section 6.1.

Proof of Lemma 3. The existence of the $(I, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget is proved by the explicit construction above. It remains to show that the number of vertices and edges in this AND gadget is $O((1/\varepsilon_1)^{c_1} I^{c_2+1})$, and the input size is $\Lambda = O((1/\varepsilon_1)^{c_1} I^{c_2})$.

By Lemma 11, the number of vertices and edges in the Type *A* and *B* AND gadgets are constants, since the parameter $\frac{1}{3}(p_2 - \varepsilon_2)$ is a constant. Let χ be a constant upper bound for these. As for Type *C* AND gadget, it has $O((1/\varepsilon)^{c_1})$ vertices and edges by Lemma 11. Since there are $2^{\log_2 I - i} \Lambda_0^{\log_2 I - i - 1} \Lambda_C$ AND gadgets in Level i and $\Lambda_C = O((1/\varepsilon)^{c_1})$ as mentioned, the total number of vertices and edges have the following bound.

$$O((1/\varepsilon)^{c_1}) + \sum_{i=1}^{\log_2 I - 1} \chi \cdot 2^{\log_2 I - i} \Lambda_0^{\log_2 I - i - 1} \Lambda_C < \chi \Lambda_C \cdot (2\Lambda_0)^{\log_2 I} = O((1/\varepsilon_1)^{c_1} I^{c_2+1}),$$

where $c_2 = \log_2 \Lambda_0$ is a constant.

As for Λ , there are $\Lambda_0^{\log_2 I - 2} \Lambda_C$ AND gadgets in each of the $\frac{I}{2}$ groups in Level 1, and each of these AND gadgets takes Λ_0 vertices as one of the two inputs. Therefore, we have

$$\Lambda = \Lambda_0 \cdot \Lambda_0^{\log_2 I - 2} \Lambda_C = O((1/\varepsilon_1)^{c_1} I^{c_2}),$$

which concludes the last part of the lemma. \square

Proof of Lemma 4. Based on Lemma 12, by changing all the Type *A* $(2, \Lambda_0, p_0, p_2, \frac{1}{3}(p_2 - \varepsilon_2), \varepsilon_2, f)$ -AND gadgets in Level 1 to the Type *A'* $(2, \Lambda_0, p^*(p_2 - \varepsilon_2), p_2, \frac{1}{3}(p_2 - \varepsilon_2), \varepsilon_2, f)$ -AND gadgets, we obtain an $(I, \Lambda, p^*(p_2 - \varepsilon_2), p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget.

The size of the AND gadget only changes by a constant, as the only difference between the two AND gadgets are the different probability scaling down gadgets used for different β for *A* and *A'*. Since the probability scaling down gadget has a constant size, we conclude the second half of the lemma. \square

6.6 Construction of Directed Edge Gadget

The $(\Upsilon, \epsilon, b, f)$ -directed edge gadget in Definition 15 can be constructed by modifying the number of layers in the $(\Lambda, p_1, p_2, \varepsilon_1, \varepsilon_2, f)$ -probability filter gadget in Definition 17. While still keeping the parameter h and α such that $a_1 h \alpha = 1 - \delta$ in the probability separation block of the probability filter gadget, we modify the number of layers in the circuit to $L = \frac{\log(\Upsilon/\epsilon)}{\log(1/(1-\delta))} + 1$.

To construct a directed edge gadget $\langle u, v \rangle$, we connect u to all the h^L inputs to the circuit, and let v be the output. The construction of directed edge gadget is shown in Figure 11.

To show property (1) in Definition 15, suppose u is connected to Υ infected vertices $v_1, v_2, \dots, v_\Upsilon$ by the directed edge gadgets. If the vertices in the i -th layer are infected with probability x_i , then

¹³If the parameter ε_2 in the $(I, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget we are constructing is not small enough to satisfy this, we can replace ε_2 with another smaller ε'_2 and instead construct a $(I, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon'_2, f)$ -AND gadget. Notice that the description 2 of Definition 13 implies that a $(I, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon'_2, f)$ -AND gadget is also a valid $(I, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget for $\varepsilon'_2 < \varepsilon_2$.

probability filter gadget with
number of layers changed to L

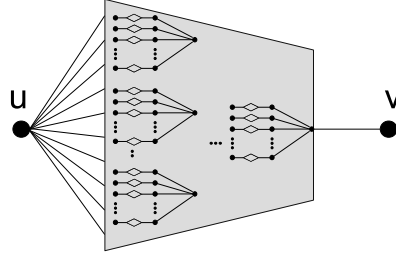


Figure 11: The directed edge gadget $\langle u, v \rangle$

the vertices in the $(i - 1)$ -th layer will be infected with probability $x_{i-1} = a_1 \alpha x_i$, which can be easily seen from Figure 7 and by noticing the symmetric property of probability scaling down gadgets mentioned in Remark 1. Therefore, each vertices in the first level that are adjacent to u will be infected with probability $(a_1 \alpha)^L$. Since there are h^L vertices in the first level and u is assumed to be connected to Υ vertices by the directed edge gadgets, the expected number of u 's infected neighbors is

$$\mathbb{E}[\text{num of infected neighbors}] = \Upsilon h^L (a_1 \alpha)^L = \Upsilon (1 - \delta)^L = \epsilon (1 - \delta) < \epsilon,$$

where recall that we have set

$$L = \frac{\log(\frac{\Upsilon}{\epsilon})}{\log \frac{1}{1-\delta}} + 1.$$

Therefore, by Markov's inequality, the probability that u has infected neighbor(s) is less than ϵ , which means u will be infected with probability less than ϵ .

For (2), suppose u is connected to v by a directed edge gadgets $\langle u, v \rangle$ and u is already infected. Then all the h^L inputs of the inner probability filter gadget will be infected with probability a_1 independently, and v will be infected with probability in $(p_2 - \epsilon_2, p_2]$ if δ is set small enough such that a_1 passes the threshold p_1 . In particular, $b > 0$.

Lastly, we prove Lemma 6 and Lemma 7.

Proof of Lemma 6. The possibility of the construction is already made explicit.

Let λ be the upper bound of the number of vertices and edges in a probability separation block in the probability filter gadget (which is a constant), the total number of vertices in a directed edge gadget is

$$\sum_{i=0}^{L-1} \lambda h^i = \lambda \frac{h^L - 1}{h - 1} = \Theta(h^L) = \Theta\left(h^{\frac{\log \Upsilon}{\log \frac{1}{1-\delta}} + \frac{\log(\frac{1}{\epsilon})}{\log \frac{1}{1-\delta}} + 1}\right) = \Theta\left(\Upsilon^d (1/\epsilon)^d\right),$$

and the total number of edges is

$$\underbrace{h^L}_{\text{number of edges from } u \text{ to the probability scaling down gadget}} + \sum_{i=0}^{L-1} \lambda h^i = \Theta(h^L) = \Theta\left(\Upsilon^d (1/\epsilon)^d\right).$$

where $d = \frac{\log h}{\log \frac{1}{1-\delta}}$.

To show that d depends only on f , it is enough to notice that we only need to setup values of h and δ such that $p_1 < a_1$ as mentioned. \square

Proof of Lemma 7. Given an $(I, \Lambda, p_0, p_2, \varepsilon_1, \varepsilon_2, f)$ -AND gadget which consists of many 2-set-input AND gadgets (see Figure 10), we can obtain a $(\Lambda, p_1, p_2, \varepsilon_1, \varepsilon_2, f)$ -probability filter gadget which is the core of an arbitrary 2-set-input AND gadget. We construct the $(\Upsilon, \epsilon, b, f)$ -directed edge gadget by increasing the number of layers in this probability filter gadget, just as what we did earlier. By our analysis above, we already have $b \in (p_2 - \varepsilon_2, p_2]$. Moreover, by Figure 8, increasing the number of layers makes b closer to p_2 . Therefore, we can have $b \in (p_2 - \frac{1}{2}\varepsilon_2, p_2]$ by just increasing the number of layers, which proves the possibility of the construction.

By our discussion in Section 6.3.2, we only need a constant number of layers to have $b \in (p_2 - \frac{1}{2}\varepsilon_2, p_2]$, as $\frac{1}{2}\varepsilon_2$ is a constant. Thus, requiring $b \in (p_2 - \frac{1}{2}\varepsilon_2, p_2]$ does not change the number of layers in an asymptotic way. Following the proof of Lemma 6, we conclude the second half of the lemma. \square

6.7 Proof of Theorem 4 for $a_1 = 0$

In the case $a_1 = 0$, the constructions of both the AND gadget and the directed edge gadget fail. Modifications of the structure in Figure 4 as well as the structure of the AND gadget are required. We will discuss these modifications in this section, and the remaining details are left to the readers.

Modification to the AND Gadget The AND gadget for the case $a_1 = 0$ is much simpler. The input $\varepsilon_1, \varepsilon_2$ is no longer needed, and both p_0, p_2 in the original AND gadget are set to $\frac{1}{2}a_2$. The definition of the modified AND gadget is shown below.

Definition 18. *A (I, Λ, f) -AND gadget takes I sets of Λ vertices each as inputs, and output a vertex such that*

1. *if the vertices in at least one input set are infected with probability 0, then the output vertex will be infected with probability 0;*
2. *if the vertices in all input sets are infected with independent probability at least $\frac{1}{2}a_2$, then the output vertex will be infected with probability at least $\frac{1}{2}a_2$,*

The construction of a $(2, \Lambda, f)$ -AND gadget is shown in Figure 12. It is easy to see that the infection of the output vertex will not affect any other vertices in this circuit due to $a_1 = 0$. Due to the same reason, property (1) above is trivial for the case $I = 2$ here. Let x be the probability that each vertex in the two input sets is infected, and let y be the probability the output is infected. Then,

$$y = \sum_{i=2}^{\Lambda} \binom{\Lambda}{i} a_i (a_2 x)^i (1 - a_2 x)^{\Lambda-i}.$$

To satisfy (2), we only need to choose Λ large enough such that $y(\frac{1}{2}a_2) \geq \frac{1}{2}a_2$. This is always possible, as we have $y(\frac{1}{2}a_2) \rightarrow p^* > \frac{1}{2}a_2$ as $\Lambda \rightarrow \infty$ (the expected number of infected neighbors of the output vertex is $\frac{1}{2}a_2\Lambda$ which goes to infinity).

Lemma 13. *For any f with $a_2 > a_1 = 0$, we can construct a $(2, \Lambda_0, f)$ -AND gadget with constant size, and Λ_0 is a constant depending on f .*

Proof. The construction above shows the existence of the gadget, and Λ_0 is a constant that is large enough to make $y(\frac{1}{2}a_2) \geq \frac{1}{2}a_2$ true, which depends only on f .

From Figure 12, it is clear that the gadget has $3\Lambda_0 + 1$ vertices and $3\Lambda_0$ edges, which are both constants. \square

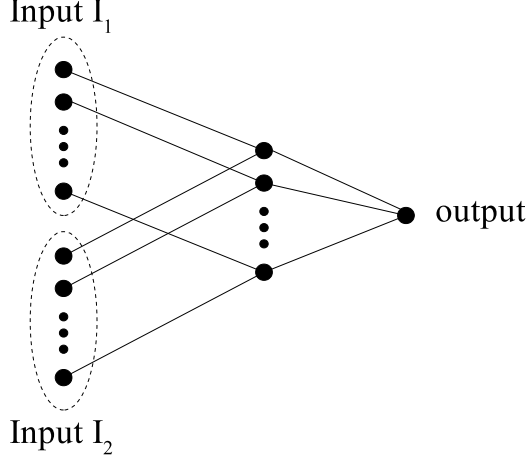


Figure 12: The modified AND gadget with parameter $(2, \Lambda, f)$

To construct a (I, Λ, f) -AND gadget with I being an integer power of 2, we use the same “tower structure” in Figure 10. Specifically, all the AND gadgets in all $\log_2 I$ levels are identically the $(2, \Lambda_0, f)$ -AND gadget in Figure 12, and the output vertices of $2^{\log_2 I - i}$ groups of $\Lambda_0^{\log_2 I - i} (2, \Lambda_0, f)$ -AND gadgets in Level i are connected to the input ends of $2^{\log_2 I - i - 1}$ groups of $\Lambda_0^{\log_2 I - i - 1} (2, \Lambda_0, f)$ -AND gadgets in Level $(i + 1)$. It is straightforward to check that (1) and (2) in Definition 18 hold for this construction.

Lemma 14. *For any f with $a_2 > a_1 = 0$ and any I that is an integer power of 2, we can construct a (I, Λ, f) -AND gadget with $O(I^{c+1})$ vertices and $O(I^{c+1})$ edges, and $\Lambda = I^c$, where c is a constant depending on f .*

Proof. The existence of this AND gadget is shown by the explicit construction.

The numbers of vertices and edges are both

$$\sum_{i=1}^{\log_2 I} 3\Lambda_0 \cdot 2^{\log_2 I - i} \Lambda_0^{\log_2 I - i} < 3\Lambda_0 \cdot (2\Lambda_0)^{\log_2 I} = O(I^{c+1}),$$

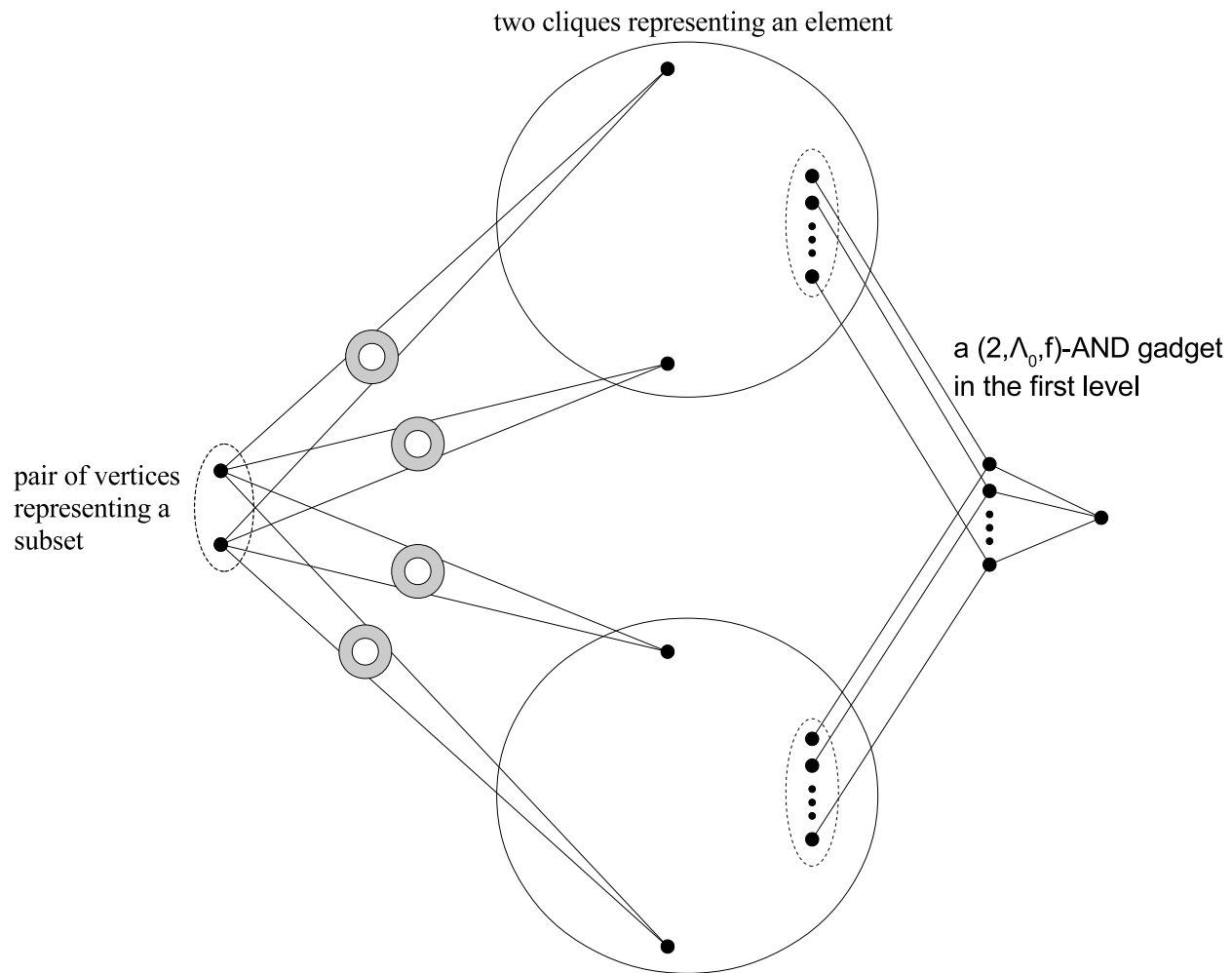
where $c = \log_2 \Lambda_0$ is a constant, and it depends only on f as Λ_0 depends only on f according to Lemma 13. Notice that the number of vertices in a $(2, \Lambda_0, f)$ -AND gadget is counted as $3\Lambda_0$ other than $3\Lambda_0 + 1$ in Lemma 13, because the output vertex of each $(2, \Lambda_0, f)$ -AND gadget is counted as one of the input vertices in one of the $(2, \Lambda_0, f)$ -AND gadgets in the next level.

Finally, since there are $\Lambda_0^{\log_2 I - 1} (2, \Lambda_0, f)$ -AND gadgets in each group in Level 1, we have

$$\Lambda = \Lambda_0 \cdot \Lambda_0^{\log_2 I - 1} = I^c,$$

which concludes the lemma. □

Modification to the Set Cover Part We will use a pair of vertices to represent a subset in the SETCOVER problem, and use a pair of cliques to represent an element in U . The pair of vertices are connected to each vertex of the two cliques by a specially designed gadget shown in the bottom of Figure 13.



The following gadget is defined:

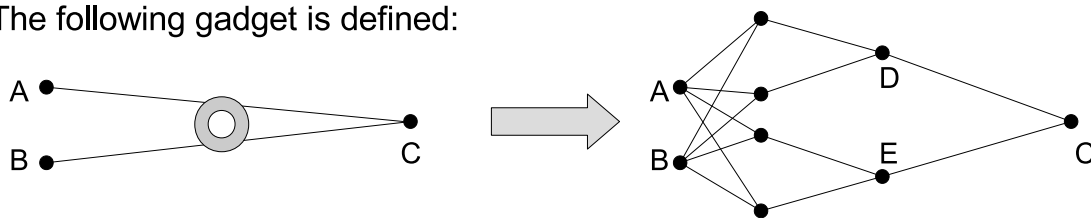


Figure 13: Connection between a pair of vertices representing a subset and vertices in the two cliques representing an element, and a $(2, \Lambda_0, f)$ -AND gadget in the first level of the $(2n, (2n)^c, f)$ -AND gadget.

If the two vertices representing a subset are both infected, it is straightforward to check that each vertex at the output end of the gadget at the bottom of Figure 13 will be infected with probability a_2^7 . Given there are m vertices in a clique, the expected number of infected vertices in a clique is $a_2^7 m$. By choosing m large enough (but still a constant) such that $a_{\lfloor a_2^7 m \rfloor} > p^* - \varepsilon$, each vertex in the clique will be infected with probability at least $p^* - \varepsilon$. Therefore, if a subset is picked such that the two vertices representing it are chosen as seeds, all pairs of cliques representing its elements will be activated. Naturally, given the SETCOVER instance in which we are choosing k subsets, we are asked to choose $2k$ seeds in the INFLUENCEMAXIMIZATION instance.

On the other hand, since $a_1 = 0$, an activated clique will not be able to infect the pair of vertices representing a subset, so the connection between the pair of vertices to each vertex in the clique is like a directed edge. Moreover, it is easy to see that we still need two seeds to pick a subset even if some cliques representing elements in this subset are activated. Although we have the option to choose the two seeds “on the gadget”, we still need to pick at least two seeds in order to “choose a subset”. Thus, it does not matter if any of these seeds is not exactly in the pair of vertices representing the subset.

The $M_2 (n, \Lambda, p^* - \varepsilon, p_2, 1/n, \varepsilon_2, f)$ -AND gadgets in Figure 4 is changed to $M_2 (2n, (2n)^c, f)$ -AND gadgets here. Moreover, each of the n groups of the $(2, \Lambda_0, f)$ -AND gadgets in Level 1 of the $(2n, (2n)^c, f)$ -AND gadget corresponds to the vertices in *the two cliques representing the same element* in U . A single $(2, \Lambda_0, f)$ -AND gadget is illustrated on the right hand side of Figure 13.

Modification to the Connection to the M_1 Vertices In Figure 4, the output vertex v is connected to the M_1 vertices by M_1 edges. Since $a_1 = 0$, such construction will fail to satisfy our purpose here. To fix this, we can use $2M_2 (2n, (2n)^c, f)$ -AND gadgets such that the outputs of every two AND gadgets are connected to each of the M_1 vertices.¹⁴

In addition, we also update the value of M_1 to $M_1 = n^{c+10}$.

Modification to the Clique Size m Since there are $(2n)^c$ vertices in each of the n inputs for each of the $2M_2 (2n, (2n)^c, f)$ -AND gadgets, to furnish enough inputs, we update the clique size to $m = 2M_2 \cdot (2n)^c = 2^{1+c} n^{2+c} = O(n^{c+2})$.

Modification to Lemma 5 To conclude this section, we have the following lemma corresponding to Lemma 5 in Section 6.1.

Lemma 15. *If the SETCOVER instance is a YES instance, by choosing $2k$ seeds appropriately, we can infect at least $\frac{1}{4} a_2^3 \cdot n^{c+12}$ vertices in expectation in the graph G we have constructed; if it is a NO instance, we can infect at most $O(kn^{c+10})$ vertices in expectation for any choice of $2k$ seeds.*

Proof. If the SETCOVER instance is a YES instance, we choose the $2k$ seeds representing the k subsets, and all the $2n$ cliques will be activated such that each vertex in all these clique will be infected with probability $p^* - \varepsilon$. Since $p^* \geq a_2$, we have $p^* - \varepsilon > \frac{1}{2} a_2$ as ε is sufficiently small due to large size of m . All the $2M_2 (2n, (2n)^c, f)$ -AND gadgets fall into case (2), so that the output vertices are infected with probabilities at least $\frac{1}{2} a_2$. The M_1 vertices in each of the $2M_2$ copies of the verification part are connected to two vertices with infection probabilities at least $\frac{1}{2} a_2$, so the

¹⁴Another way to fix this is to reduce the number of levels by 1 in the $(2n, (2n)^c, f)$ -AND gadget, such that we have two output vertices of the AND gadget instead of only one output in Definition 18.

total expected number of infected vertices in G is at least

$$M_2 \times \underbrace{\left(\frac{1}{2}a_2\right)^2}_{\text{probability that both output vertices are infected}} \times \underbrace{(a_2 M_1)}_{\text{expected num of infections in } M_1 \text{ vertices}} = \frac{1}{4}a_2^3 \cdot n^{c+12}.$$

If the SETCOVER instance is a NO instance, consider any choice of $2k$ seeds with k_1 seeds in the vertices representing subsets, k_2 seeds in the connection gadgets between vertices representing subsets and vertices in the cliques, k_3 seeds in the $2n$ cliques, k_4 seeds in those $(2, \Lambda_0, f)$ -AND gadgets at Level 1 of the $(2n, (2n)^c, f)$ -AND gadgets, and $k_5 = 2k - k_1 - k_2 - k_3 - k_4$ seeds in the remaining part of the verification parts (the $(2, \Lambda_0, f)$ -AND gadgets at the remaining levels and the M_1 vertices connecting to the $(2n, (2n)^c, f)$ -AND gadgets). Again, we first prove that at least one clique will not be activated such that all its vertices are infected with probability 0.

First of all, those k_5 seeds cannot have effect in activating cliques. This is because their influence cannot pass through the $(2, \Lambda_0, f)$ -AND gadgets in the first level, as the infection of the output vertex in each $(2, \Lambda_0, f)$ -AND gadget cannot further infect the input vertices due to $a_1 = 0$.

Secondly, for those k_1 and k_2 seeds, they are at the vertex-pairs representing the subsets and the gadgets connected to those pairs respectively. We call the vertices in those gadgets connecting to a pair *the vertices around the pair*. It is easy to see that we need to choose at least 2 seeds in or around a pair, in order to pick a subset. To see this, even if vertex C in the gadget (at the bottom of Figure 13) is already infected (which is possible as C belongs to a clique which may have been activated already) such that D and E already have one infected neighbor, we still cannot make both A and B infected by picking only 1 seed in or around the pair (A, B) . Thus, we assume without loss of generality that all k_2 seeds are on the pairs representing the subsets, as we need at least 2 seeds in or around a pair (A, B) in which case we can assume the seeds are just at A and B .

For those k_3 seeds on the cliques and k_4 seeds on the $(2, \Lambda_0, f)$ -AND gadgets in the first level, since each AND gadget in the first level takes two sets of vertices from two cliques *representing the same element in U* , we need at least 3 seeds to activate two cliques representing the same element in U : one in the middle of the AND gadget, and one in each of the two cliques (such that the two vertices connecting to the seed in the middle of the AND gadget have two infected neighbors, and stand a chance to activate the two cliques). In contrast, we only need 2 seeds to activate these two cliques, by choosing the pair of vertices representing the subset covering the element that these two cliques represent. Therefore, we can assume that those k_3 and k_4 seeds are also on those pairs representing subsets.

Since $k_1 + k_2 + k_3 + k_4 \leq 2k$ and the SETCOVER instance is a NO instance, by the fact that we need 2 seeds to pick a subset, we conclude that at least one clique will not be activated, and the vertices in this clique are infected with probability 0.

By the effect of the $(2n, (2n)^c, f)$ -AND gadget, except for those (at most) $k_4 + k_5$ AND gadgets containing seeds, the output vertices of the remaining $2M_2 - k_4 - k_5$ AND gadgets will be infected with probability 0, which have no effect on those M_1 vertices. Therefore, even if all the vertices in the set cover part, the $k_4 + k_5$ copies of the verification parts, and the $2M_2$ $(2n, (2n)^c, f)$ -AND gadgets are infected, the total number of infected vertices cannot exceeds

$$\underbrace{2\kappa + 6\kappa(2n)m + 2nm + 2M_2}_{\text{size of the set cover part}} \underbrace{(2n)^{c+1}}_{\text{size of an AND gadget}} + (k_4 + k_5) \underbrace{((2n)^{c+1} + M_1)}_{\text{size of a verification part}} = O(kn^{c+10}),$$

which concludes the lemma. \square

Noticing that the total number of vertices in G is

$$N = 2\kappa + 6\kappa(2n)m + 2nm + M_2((2n)^{c+1} + M_1) = O(n^{c+12}),$$

and

$$\frac{\frac{1}{4}a_2^3 \cdot n^{c+12}}{\Omega(kn^{c+10})} = O(n),$$

we conclude Theorem 4 in the case $a_1 = 0$ by setting $\tau = \frac{1}{c+12}$.

7 Conclusion

We show the hardness of approximating INFMAX in several settings restricting the network structure or the cascade model. Before our results there was some hope that the hardness of nonsubmodular influence maximization was only caused by the hardness of detecting community structure within the network. However, our results show that even for very plain community structures, INFMAX can remain hard. We show that it is the bidirectional nature of contagions which renders the problem hard.

We also show the inapproximability of INFMAX even in the restrictive universal local influence model (Definition 9) with 2-submodular local influence function f which is almost submodular. Since it turns out assumptions on either graph topology or cascade model do not really make INFMAX easy, a natural question is that, what if we make assumptions on both?

We conclude with the following open problem: considering the universal local influence model in Definition 9 with 2-submodular f on the stochastic hierarchical blockmodel¹⁵, does there exist a 2-submodular f , such that INFMAX is NP-hard to approximate to within a constant factor? Or is it the case that for any 2-submodular f , there exists a constant factor approximation to INFMAX?

¹⁵It does not make much sense to consider universal local influence model I_f^G on the hierarchical blockmodel, as f depends only on the *number* of infected neighbors and ignores the weights of edges connecting to the neighbors, while the hierarchical blockmodel considers weighted graphs.

References

- [1] Rico Angell and Grant Schoenebeck. Don't be greedy: Leveraging community structure to find high quality seed sets for influence maximization. *arXiv*, 2016.
- [2] W. B. Arthur. Competing technologies, increasing returns, and lock-in by historical events. *Economic Journal*, 99(394):pp. 116–131, 1989.
- [3] Lars Backstrom, Daniel P. Huttenlocher, Jon M. Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *ACM SIGKDD*, 2006.
- [4] Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. The diffusion of microfinance. *Science*, 341(6144), 2013.
- [5] F. M. Bass. A new product growth for model consumer durables. *Management Science*, 15(5):215–227, 1969.
- [6] Shishir Bharathi, David Kempe, and Mahyar Salek. *WINE*, chapter Competitive Influence Maximization in Social Networks, pages 306–311. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.
- [7] Christian Borgs, Michael Brautbar, Jennifer T Chayes, and Brendan Lucier. Influence maximization in social networks: Towards an optimal algorithmic solution. *CoRR*, 2012.
- [8] J. J. Brown and P. H. Reingen. Social ties and word-of-mouth referral behavior. *Journal of Consumer Research*, 14:350–362, 1987.
- [9] D. Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 2010.
- [10] Ning Chen. On the approximability of influence in social networks. *SIAM Journal on Discrete Mathematics*, 23(3):1400–1415, 2009.
- [11] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *ACM SIGKDD*, pages 199–208. ACM, 2009.
- [12] Wei Chen, Yifei Yuan, and Li Zhang. Scalable influence maximization in social networks under the linear threshold model. In *2010 IEEE International Conference on Data Mining*, pages 88–97. IEEE, 2010.
- [13] Aaron Clauset, Cristopher Moore, and Mark EJ Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.
- [14] J. Coleman, E. Katz, and H. Menzel. The diffusion of an innovation among physicians. *Sociometry*, 20:253–270, 1957.
- [15] T. G. Conley and C. R. Udry. Learning about a new technology: Pineapple in Ghana. *American Economic Review*, 100(1):35–69, 2010.
- [16] Paul DiMaggio. Structural analysis of organizational fields: A blockmodel approach. *Research in organizational behavior*, 1986.

- [17] P. Domingos and M. Richardson. Mining the network value of customers. In *ACM SIGKDD*, 2001.
- [18] J. Goldenberg, B. Libai, and E. Muller. Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. *Academy of Marketing Science Review*, 9(3):1–18, 2001.
- [19] Sanjeev Goyal and Michael Kearns. Competitive contagion in networks. In *STOC*, pages 759–774, 2012.
- [20] Mark Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.
- [21] Xinran He and David Kempe. Robust influence maximization. In *ACM SIGKDD*, 2016.
- [22] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.
- [23] Matthew O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008.
- [24] David Kempe, Jon M. Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *ACM SIGKDD*, pages 137–146, 2003.
- [25] David Kempe, Jon M. Kleinberg, and Éva Tardos. Influential nodes in a diffusion model for social networks. In *ICALP*, pages 1127–1138, 2005.
- [26] K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In *ICWSM*, pages 90–97, 2010.
- [27] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. In *EC*, pages 228–237, 2006.
- [28] Brendan Lucier, Joel Oren, and Yaron Singer. Influence at scale: Distributed computation of complex contagion in networks. In *ACM SIGKDD*, pages 735–744. ACM, 2015.
- [29] Vince Lyzinski, Minh Tang, Avanti Athreya, Youngser Park, and Carey E Priebe. Community detection and classification in hierarchical stochastic blockmodels. *arXiv*, 2015.
- [30] V. Mahajan, E. Muller, and F. M. Bass. New product diffusion models in marketing: A review and directions for research. *Journal of Marketing*, 54:1–26, 1990.
- [31] S. Morris. Contagion. *Review of Economic Studies*, 67(1):57–78, 2000.
- [32] Elchanan Mossel and Sébastien Roch. Submodularity of influence in social networks: From local to global. *SIAM J. Comput.*, 39(6):2176–2188, 2010.
- [33] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.
- [34] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *ACM SIGKDD*, pages 61–70, 2002.

- [35] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics : Idioms , political hashtags , and complex contagion on twitter. In *WWW*, pages 695–704. ACM, 2011.
- [36] Lior Seeman and Yaron Singer. Adaptive seeding in social networks. In *FOCS*, pages 459–468. IEEE, 2013.
- [37] Yaron Singer Thibaut Horel. Maximization of approximately submodular functions. In *NIPS*, 2016.
- [38] D. J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.
- [39] Harrison C White, Scott A Boorman, and Ronald L Breiger. Social structure from multiple networks. i. blockmodels of roles and positions. *American Journal of Sociology*, 81(4):730–780, 1976.