Influence Maximization on Undirected Graphs: Towards Closing the $(1 - 1/e)$ Gap

GRANT SCHOENEBECK, The University of Michigan, USA

BIAOSHUAI TAO, The University of Michigan, USA

We study the influence maximization problem in undirected networks, specifically focusing on the independent cascade and linear threshold models. We prove APX-hardness (NP-hardness of approximation within factor $(1 - \tau)$ for some constant $\tau > 0$) for both models, which improves the previous NP-hardness lower bound for the linear threshold model. No previous hardness result was known for the independent cascade model.

As part of the hardness proof, we show some natural properties of these cascades on undirected graphs. For example, we show that the expected number of infections of a seed set S is upper-bounded by the size of the edge cut of S in the linear threshold model and a special case of the independent cascade model called the weighted independent cascade model.

Motivated by our upper bounds, we present a suite of highly scalable local greedy heuristics for the influence maximization problem on both the linear threshold model and the weighted independent cascade model on undirected graphs that, in practice, find seed sets which on average obtain 97.52% of the performance of the much slower greedy algorithm for the linear threshold model, and 97.39% of the performance of the greedy algorithm for the weighted independent cascade model. Our heuristics also outperform other popular local heuristics, such as the degree discount heuristic by Chen et al. [6].

CCS Concepts: • **Networks** → **Network economics**; • **Theory of computation** → **Social networks**; *Problems, reductions and completeness*; Approximation algorithms analysis.

Additional Key Words and Phrases: influence maximization; linear threshold model; independent cascade model; PCP theorem

ACM Reference Format:

Grant Schoenebeck and Biaoshuai Tao. 2019. Influence Maximization on Undirected Graphs: Towards Closing the $(1 - 1/e)$ Gap. In *ACM EC '19: ACM Conference on Economics and Computation (EC '19), June 24–28, 2019, Phoenix, AZ, USA*. ACM, New York, NY, USA, 31 pages. <https://doi.org/10.1145/3328526.3329650>

1 INTRODUCTION

In social networks, *cascades* model the phenomena that agents receive certain information, adopt certain products, or take up certain political opinions due to the influence from their neighbors. A natural and well-studied problem is, given a network, which agents to initially seed in order to maximize the number of infections of the resulting cascade. A vast array of works test heuristics for this problem, usually on undirected graphs. However, most theoretical literature has previously ignored undirected graphs. The algorithms do not take advantage of the undirected nature of the

The authors gratefully acknowledge the support of the National Science Foundation under Career Award 1452915 and AifT Award 1535912.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

EC '19, June 24–28, 2019, Phoenix, AZ, USA

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6792-9/19/06...\$15.00

<https://doi.org/10.1145/3328526.3329650>

problems, while the lower bounds (especially inapproximability results) require directed networks to exhibit worst-case hardness.

In the *influence maximization problem* (INFMAX), a central planner is given a graph and a limited budget k , and he needs to pick k seeds such that the expected total number of infected vertices in the graph at the end of the cascade is maximized. This problem plays a central role in viral marketing—a marketing strategy of advertising products by giving the products to a certain number of users for free in the hope that they will recommend these products to their friends.

Perhaps the two best known cascade models are the *independent cascade model* and the *linear threshold model*, both introduced in the seminal paper Kempe et al. [17]. These two models are studied almost exclusively in the past literature of this field, and they will also be the focus of this paper. In the independent cascade model, a newly-infected vertex (or seed) u infects each of its not-yet-infected neighbors v with a fixed probability independently. (See Section 2.1 for more details.) In the linear threshold model, in the case the graph is unweighted, each non-seed vertex has a threshold sampled uniformly and independently from the interval $[0, 1]$, and becomes infected when the fraction of its infected neighbors exceeds this threshold. (See Section 2.2 for more details.)

Both models were shown to be *submodular*¹ even in the directed, weighted cases [17], which implies that the greedy algorithm which iteratively picks the seed that has the most marginal influence achieve a $(1 - 1/e)$ -approximation. On the hardness side, it is shown by Kempe et al. [17] that INFMAX is NP-hard to approximate to within a factor of $(1 - 1/e + \varepsilon)$ for any ε for the independent cascade model², and finding the exact optimal solution for INFMAX for the linear threshold model is NP-hard, even for undirected graphs³.

The greedy algorithm, unfortunately, is widely known to be limited in its scalability (see, e.g., [2]). As a results, new INFMAX algorithms/heuristics are invented almost every year after [17], and they are tested on real social networks in our daily life (see Sect. 1.2 for details). Among these social networks on which INFMAX algorithms are tested, most of them are undirected. For example, Arora et al. [2] performs a thoughtful benchmarking study of those INFMAX algorithms, and the four main graphs considered, Nethept, HepPh, DBLP, YouTube, are all undirected. Undirected datasets are used in most, if not all, of the empirical work in Sect. 1.2, and are used exclusively in [6, 7, 13–15]⁴ in particular. Peer relationships are, in some sense, undirected by definition. Moreover, these models typically try to capture word-of-mouth influence among peers as opposed to a more directional advertisements.

Unfortunately, theoretical results for INFMAX on undirected graphs are lacking. The original greedy algorithm in [17], for example, does not take advantage of the undirected nature of the graph in any meaningful way. On the positive side, for the independent cascade model with undirected graphs, Khanna and Lucier [19] improve the analysis of the greedy algorithm to show it obtains an approximation ratio slightly better than $(1 - 1/e)$. This is one of the very few existing results that take advantage of the undirected nature of the graph. On the hardness side, the inapproximability results for INFMAX on undirected graphs are almost completely lacking in the literature. For the linear threshold model, the previously mentioned NP-hardness result in [17], which is almost 16

¹A cascade model is *submodular* if the marginal influence from a newly infected vertex u to its not-yet-infected neighbor v is decreasing when v is having increasing number of existing infected neighbors. More formally, letting T be the set of v 's infected neighbors and letting $P(T)$ be the probability that v will be infected given the set of infected neighbors T , a cascade model is submodular if $P(T_1 \cup \{v\}) - P(T_1) \geq P(T_2 \cup \{v\}) - P(T_2)$ whenever $T_1 \subseteq T_2$ and $v \notin T_2$.

²Although not being explicitly stated in [17], the proof for the NP-hardness considers a reduction from the maximum-coverage problem, which straightforwardly implies a factor $(1 - 1/e + \varepsilon)$ hardness-of-approximation.

³Again, it is not explicitly stated in [17] that the NP-hardness holds for undirected graphs, but the reduction in [17] uses an undirected graph.

⁴Directed graphs are treated as undirected graphs in [7].

years old, is the only known hardness result to the best of our knowledge. There is still a gap between $(1 - 1/e)$ and 1 for both directed and undirected graphs. For the independent cascade model, although the inapproximability factor $(1 - 1/e + \epsilon)$ in [17] matches the approximation guarantee $(1 - 1/e)$ of the greedy algorithm, the hardness/inapproximability results for undirected graphs, to the best of our knowledge, are completely missing.⁵ In particular, before this paper, it was not even known whether computing the exact optimal seeds is NP-hard.

It is worth noticing that, given a seed set S , computing the exact expected number of infected vertices is #P-hard [8]. This can be easily extended to the #P-hardness of INFMAX. However, if we assume oracle access to the exact expected number of infected vertices given seed set S , a setting commonly adopted in most work (including [17]), the complexity of computing the exact optimal seed set is still unknown. Moreover, the expected number of infected vertices can be estimated by a Markov chain Monte Carlo (MCMC) simulation, and it can be approximated arbitrarily close by increasing the number of MCMC simulations (typically, the number of simulations is set to 10,000 in practice, as suggested in [17]). Therefore, the #P-hardness result cannot even rule out a polynomial time approximation scheme (PTAS) for INFMAX. In this paper, we will assume that the expected number of infected neighbors can be accessed by an oracle.

1.1 Our Results

As our main results, we show that INFMAX on both the independent cascade model and the linear threshold model with undirected graphs is APX-hard: there exists a constant $\tau > 0$ such that approximating INFMAX to within factor $(1 - \tau)$ is NP-hard. The APX-hardness holds even for the two well-studied special cases of the independent cascade model: *the uniform independent cascade model* and *the weighted independent cascade model*.

We prove this result by using the PCP theorems and a novel coupling approach. Our coupling approach reveals an upper bound on the expected number of infected vertices for each of the three cascade models: the linear threshold model, the weighted independent cascade model and the uniform independent cascade model. An especially interesting upper bound for both the linear threshold model and the weighted independent cascade model on undirected graphs is that, *the expected number of vertices infected by a seed set S is no more than the number of edges between vertices in S and vertices outside S* . This is an indication that the cascade on these two models on undirected graphs is diminishing and somehow characterized by its local behaviors on S and neighbors of vertices in S : no matter how large and dense the graph G is outside S , the cascade is limited as long as S has limited connections to the remaining vertices. This indicates that, although the weighted independent cascade model is similar to the uniform independent cascade model in definition, it is more similar to the linear threshold model in behaviors.⁶ This also theoretically justifies the phenomenon that those reverse-reachable-set-based algorithms are much more scalable on the linear threshold model and the weighted independent cascade model than the uniform independent cascade model, as observed by Arora et al. [2] (See Sect. 4.3 for details.).

Motivated by this observation, we present a family of INFMAX heuristics we called *the local greedy heuristics*, which iteratively select the seeds only based on the local features of S : the number of edges goes out from S , the number of vertices that are connected to S , and a combination of both. Our heuristics are highly scalable, since we have only checked the neighbors of the seeds, instead of analyzing the influence of the seeds in the graph globally (as it is in the MCMC simulations). Moreover, our heuristics, although being extremely simple, produce seeds with almost the same

⁵An APX-hardness result for the independent cascade INFMAX with undirected graphs was presented by Khanna and Lucier [19]. However, their APX-hardness result is applied to a more general non-standard setting. We discussed this in details in Sect. 1.2.

⁶Similar observations have been made before by Chen et al. [6], Kempe et al. [17].

quality as the seeds output by those greedy-based algorithms in the state-of-art (being 97.52% of the greedy-based algorithms on average for the linear threshold model, and 97.39% for the weighted independent cascade model), outperform *the degree discount heuristic* proposed by [6] which iteratively selects a vertex with highest degree and removes it from the graph, and significantly outperform the naïve algorithm that picks the seeds with the highest degrees.

1.2 Related work

Theoretical work. The influence maximization problem was initially posed by Domingos and Richardson [11, 27]. The linear threshold model and the independent cascade model were introduced by Kempe, Kleinberg, and Tardos [17]. Kempe et al. [17] showed that the simple greedy algorithm that iteratively chooses the seed with the maximum marginal influence obtains a $(1 - 1/e)$ factor approximation to the problem in both models. The same set of authors later extended this result to a family of submodular cascades, called the decreasing cascade model, which captures the linear threshold model and the independent cascade model as special cases [18]. Mossel and Roch [24] showed that local submodularity implies global submodularity: if the cascade model is submodular, then the global influence function that maps a set of seeds to the expected total number of infections is a submodular function. This concludes that the greedy algorithm achieve a $(1 - 1/e)$ factor approximation whenever the cascade model is submodular, as the greedy algorithm achieves $(1 - 1/e)$ approximation whenever the objective function is monotone and submodular, which is known by Nemhauser et al. [25] and also remarked by Kempe et al. [17].

On the hardness or inapproximability side, Kempe et al. [17] showed that INFMAX on both the linear threshold model and the independent cascade model is NP-hard. For the independent cascade model, Kempe et al. [17] considered a reduction from the maximum coverage problem preserving the approximation factor. Since Feige [12] showed that the maximum coverage problem is NP-hard to approximated within factor $(1 - 1/e + \epsilon)$ for any $\epsilon > 0$, the same inapproximability factor holds for the independent cascade INFMAX. Therefore, the gap between upper bound and lower bound for the independent cascade INFMAX is closed. Note that, however, the reduction depends heavily on the directed nature of the graph, and there is no known hardness result for the independent cascade INFMAX with undirected graphs. For the linear threshold model, Kempe et al. [17] reduced the problem from the vertex cover problem such that the graph, being undirected, is the same in the two problems. It is easy to see that we can infect all the vertices in the graph with probability 1 by k seeds if and only if the graph admits a vertex cover of size k (the k seeds are exactly the vertex cover). This reduction shows the linear threshold model is NP-hard even for undirected graphs, but does not imply any hardness of approximation.

If the cascade model can be nonsubmodular, Kempe et al. [17] showed that INFMAX is NP-hard to approximate to within a factor of $N^{1-\epsilon}$ for any $\epsilon > 0$. Many work after this [4, 22, 28, 29, 32] are dedicated to study specific nonsubmodular models (or study optimization on nonsubmodular objective functions), and unfortunately most of those models, even for those very restrictive models, admit polynomial inapproximability factors ($N^{1-\tau}$ for some constant $\tau > 0$, or even worse, $N^{1-\epsilon}$ for any $\epsilon > 0$).

Most relevant to our work, Khanna and Lucier [19] studied the independent cascade INFMAX on undirected graphs. They worked on a more general model where the seed-picker is only allowed to pick seeds from a prescribed subset of vertices and the vertices are weighted so that the objective is to maximize the expected total weight of all the infected vertices. Under this generalized setting, Khanna and Lucier [19] showed that the greedy algorithm achieves a $(1 - 1/e + c)$ -approximation for some constant $c > 0$, and INFMAX on this model is APX-hard. In their APX-hardness reduction, a bipartite graph is constructed such that the vertices on the left-hand side have weight 0 and the vertices on the right-hand side have weight 1, and the seed-picker is restricted to pick seeds

among the left-hand side vertices. These features are not allowed in the standard INFMAX setting (where the seed-picker can seed on any vertices, and the expected *number* of infected vertices is the objective), and extending the same APX-hardness result to the less general, standard setting requires significantly more insight.

A weaker version of our upper bound for the linear threshold model on undirected graphs was presented by Lim et al. [23], where it was shown that the expected number of infected vertices by a single seed is upper-bounded by the degree of this seed plus one. In addition, Lim et al. [23] observed that this upper bound is tight when the graph is a tree. We have the same observation in Lemma 4.2 and Corollary 4.4. Our upper bound generalizes Lim et al.'s upper bound to the setting with multiple seeds, which is crucial for showing the APX-hardness of the INFMAX problem.

Empirical work. Following the work of Kempe, Kleinberg, and Tardos [17, 18], there was extensive work to solve INFMAX based on the alternative implementations of the greedy algorithm designed to be efficient and scalable. Leskovec et al. [20] and Goyal et al. [14] propose algorithms CELF and CELF++ respectively which improve the greedy algorithm by skipping some of the vertices in each iteration that are known to be suboptimal, but still relying on MCMC simulations to evaluate the influence of a given seed set. Although the $(1 - 1/e)$ approximation guarantee is secured, algorithms based on MCMC simulations are limited in their scalability [2].

To avoid using MCMC simulations to evaluate the influence of a given seed set, another category of algorithms are based on constructing *reverse reachable sets*, these includes RIS [3], TIM⁺ [31] and IMM [30]. A reverse reachable set is obtained by selecting a vertex uniformly at random, and sampling the live edges to obtain a set of vertices infecting which can cause the infection of the selected vertex. In these algorithms, a sufficient number of reverse reachable sets are sampled, and a greedy maximum coverage algorithm is implemented to select the seeds (more details are discussed in Sect. 4.3). Theoretically, these algorithms achieve $(1 - 1/e - \varepsilon)$ approximation, and arbitrarily small $\varepsilon > 0$ can be obtained by sampling more reverse reachable sets.⁷

Another method similar to reverse reachable sets sampling is *snapshots* sampling, based on which the algorithms StaticGreedy [9] and PMC [26] were invented. Instead of considering the set of vertices that can infect a particular vertex, these algorithms sample a sufficient number of “snapshots” of the graph, each of which is obtained by including each edge with probability proportional to its edge-weight.

Other than the above-mentioned algorithms that preserve the theoretical approximation guarantee, heuristics have also been proposed, and most of them are based on computing a “score” for a seed set that are closely correlated to the expected number of infected vertices. These algorithms include the Degree Discount Heuristics [6], PMIA [8], LDAG [7], SIMPATH [15], IRIE [16] and EaSyIM [13]. Our heuristic in Sect. 5 is perhaps most related to the Degree Discount Heuristics, as both heuristics only examine the neighbors or a restrictive local area around the seeds. As we will see in Sect. 5, by exploiting our theoretically provable upper bounds in Sect. 4, our heuristics produce seeds with better qualities.

A benchmarking study of most of these above-mentioned algorithms has been done by Arora et al. [2], and it was observed that TIM⁺, IMM and PMC provide the seeds with the best qualities while also maintain moderate scalability. As a result, we will mainly compare the performance of our heuristics to those reverse-reachable-set-based algorithms.

Rather than working on the linear threshold model and the independent cascade model as it is in all the work above, empirical study for nonsubmodular INFMAX has also been studied. For example,

⁷There was a mistake in IMM which makes IMM possibly fail to achieve the said approximation guarantee. Chen [5] spotted and corrected this mistake.

Angell and Schoenebeck [1] provided a dynamic programming based INFMAX algorithm which is fundamentally different than most algorithms that are more or less based on greedy.

2 PRELIMINARIES

We consider the two mostly studied cascade models, *the independent cascade model* and *the linear threshold model*, introduced in [17].

Throughout this paper, we use $G = (V, E, w)$ to represent the weighted graph which may or may not be directed. We use S to denote the set of seeds. Denote $k = |S|$ and $N = |V|$, and denote by $\deg(v)$ the degree of v when G is undirected and the *in-degree* of vertex v otherwise. For each $v \in V$, let $\Gamma(v) = \{u : (u, v) \in E\}$ be the set of *in-neighbors* of vertex v , and $\Gamma(v)$ is just the set of neighbors if G is undirected.

2.1 Independent Cascade Model

Definition 2.1. The *independent cascade model* IC_G is defined by a directed edge-weighted graph $G = (V, E, w)$ such that $w(u, v) \leq 1$ for each $(u, v) \in E$. On input seed set $S \subseteq V$, $IC_G(S)$ outputs a set of infected vertices as follows:

- (1) Initially at iteration 0, only vertices in S are infected.
- (2) In each iteration $i = 1, 2, \dots$, each vertex u infected in the previous iteration, $i - 1$, infects each (not yet infected) out-neighbor v with probability $w(u, v)$.
- (3) After an iteration where there is no additional infected vertices, $IC_G(S)$ outputs the set of infected vertices, as no additional infections can occur.

The independent cascade model can be viewed in the following equivalent way, known as *the live-edge interpretation*.

THEOREM 2.2 (CLAIM 2.3 IN [17]). Let $\widehat{IC}_G(S) \subseteq V$ be the set of vertices that are reachable from S when each edge (u, v) is included in the graph with an independent probability $w(u, v)$. Then $\widehat{IC}_G(S)$ and $IC_G(S)$ have the same distribution.

It is easy to see the equivalence of both descriptions, instead of flipping a coin with probability $w(u, v)$ to decide if u successfully infects v , we can flip all the $|E|$ coins at the beginning of the cascade.

Definition 2.3. The *uniform independent cascade model* UIC_G^p for $G = (V, E)$ is a special case of the independent cascade model IC_G with $G = (V, E, w)$, where $w(u, v) = p$ for each $(u, v) \in E$ and $p \in [0, 1]$ is an input parameter.

Definition 2.4. The *weighted independent cascade model* WIC_G for $G = (V, E)$ is a special case of the independent cascade model IC_G with $G = (V, E, w)$, where $w(u, v) = 1/\deg(v)$ for each $(u, v) \in E$.

Bidirecting an undirected graph. When we restrict our attention for undirected graphs, the undirected graph is “bidirected”, i.e., it is viewed as a special directed graph with each undirected edge of the graph being viewed as two anti-parallel edges.

It should then be noticed that modeling cascade process on undirected graphs does not necessarily imply that the influence from u to v is the same as the influence from v to u on an edge (u, v) . Although they are the same in the uniform independent cascade model, they are not necessarily the same in the weighted independent cascade model. In the weighted independent cascade model, if u has a larger degree than v , it is less likely that v will infect u than u will infect v , as $w(v, u) = \frac{1}{\deg(u)} < \frac{1}{\deg(v)} = w(u, v)$.

Due to space constraints, much of the material relating to the independent cascade model and its variants is deferred to the appendix, and instead the main body focuses on the linear threshold model.

2.2 Linear Threshold Model

Definition 2.5. The *linear threshold model* LT_G is defined by a directed edge-weighted graph $G = (V, E, w)$ such that $\sum_{u:u \in \Gamma(v)} w(u, v) \leq 1$ for each $v \in V$. On input seed set $S \subseteq V$, $LT_G(S)$ outputs a set of infected vertices as follows:

- (1) Initially, only vertices in S are infected, and for each vertex v a *threshold* θ_v is sampled uniformly at random from $[0, 1]$ independently.
- (2) In each subsequent iteration, a vertex v becomes infected if the sum of the weights of edges connecting v 's infected in-neighbors to v exceeds the threshold θ_v :

$$\sum_{u:u \in \Gamma(v) \text{ and } u \text{ is infected}} w(u, v) \geq \theta_v.$$

- (3) After an iteration where there is no additional infected vertices, $LT_G(S)$ outputs the set of infected vertices.

The linear threshold model also has its live-edge interpretation, although the interpretation is less intuitive compared to the interpretation of the independent cascade model.

THEOREM 2.6 (CLAIM 2.6 IN [17]). Let $\widehat{LT}_G(S) \subseteq V$ be the set of vertices that are reachable from S when each vertex v picks at most one of its incoming edge to be included in the graph such that with probability $w(u, v)$ the incoming edge (u, v) is picked and with probability $1 - \sum_{u:u \in \Gamma(v)} w(u, v)$ no incoming edge is picked⁸, and vertices pick their incoming edges independently. Then $\widehat{LT}_G(S)$ and $LT_G(S)$ have the same distribution.

For an intuition of this interpretation, consider a not-yet-infected vertex v and a set of its infected neighbors $IN(v) \subseteq \Gamma(v)$. v will be infected by vertices in $IN(v)$ with probability $\sum_{u:u \in IN(v)} w(u, v)$, as $\Pr(\theta_v \leq \sum_{u:u \in IN(v)} w(u, v)) = \sum_{u:u \in IN(v)} w(u, v)$. In the case where v becomes infected, we can attribute its infection to exactly one of its infected neighbors. The infection will be attributed to neighboring infected vertex u with probability equal to $w(u, v)$. Overall, the probability that v includes an incoming edge from $\{(u, v) : u \in IN(v)\}$ is exactly $\sum_{u:u \in IN(v)} w(u, v)$.

Definition 2.7. The *uniform linear threshold model* ULT_G for $G = (V, E)$ is a special case of the linear threshold model LT_G with $G = (V, E, w)$, where $w(u, v) = 1/\deg(v)$ for each $(u, v) \in E$.

Again, when we restrict our attention for undirected graphs, the undirected graph is “bidirected”. In the case of the linear threshold model, the uniform linear threshold model is automatically considered if G is undirected. An undirected edge (u, v) is viewed as two directed edges (u, v) and (v, u) , with $w(u, v) = 1/\deg(v)$ and $w(v, u) = 1/\deg(u)$. The same as it is in the weighted independent cascade model, if u has a larger degree than v , it is less likely that v will infect u than u will infect v .

2.3 Influence Maximization Problem

Let $\sigma_G(S)$ be the *expected* total number of infected vertices due to the influence of S . In the linear threshold model, $\sigma_G(S) = \mathbb{E}[|LT_G(S)|]$, where the expectation is taken over the samplings of thresholds of all vertices; in the independent cascade model, $\sigma_G(S) = \mathbb{E}[|IC_G(S)|]$, where the expectation is taken over the samplings of all edges. In the INFMAX problem, one is given as

⁸More formally, v 's in-neighbors are ordered $u_1, \dots, u_T$ and a real number r in $[0, 1]$ is sampled uniformly, the t -th edge (u_t, v) is picked if $r \in [\sum_{i=1}^{t-1} w(u_i, v), \sum_{i=1}^t w(u_i, v))$, and no edge is picked if $r \geq \sum_{i=1}^T w(u_i, v)$.

inputs an edge-weighted graph $G = (V, E, w)$ and an integer k , and finds a seed set S of size k that maximizes the global influence $\sigma_G(S)$. As we have mentioned in the introduction, we assume oracle access to $\sigma_G(\cdot)$.

2.4 Dinur's PCP Theorem

We will use the following theorem, which is a corollary to a version of the PCP theorem due to Dinur [10]. In Appendix A, we review Dinur's PCP theorem and how it implies Theorem 2.8.

THEOREM 2.8 (GAP OF INDSET). *There exist a universal constant integer d and a universal constant $\gamma \in (0, 1)$ such that, given an undirected graph $G = (V, E)$ where the degree of each vertex is at most d and the number of vertices $|V|$ is a multiple of 3, it is NP-hard to distinguish between the following two cases:*

- YES: G has an independent set of size n ;
- NO: any subgraph of G induced by n vertices contains at least γn edges,

where $n = \frac{|V|}{3}$.

3 APX-HARDNESS OF INFMAX

Our APX-hardness result for the linear threshold model is built upon the upper bound on $\sigma_G(S)$ in Theorem 3.1. The upper bound is particularly interesting even on its own, and it will further be considered in Sect. 5. It shows that the expected number of vertices infected by S is upper-bounded by the number of edges between S and $V \setminus S$, which reveals that $\sigma_G(S)$ is somehow characterized by its local behaviors on S and neighbors of vertices in S . In particular, the expected number of new infections caused by a single seed is upper-bounded by the degree of this seed. As we will see later in Sect. 4.1, the expected number of infections caused by a single seed is exactly the degree of this seed if the graph is a tree (Lemma 4.2). This implies that trees provide the most number of infections, and adding more edges to a tree may only reduce the total number of infections.

THEOREM 3.1. *Given a uniform linear threshold model ULT_G with an undirected graph $G = (V, E)$ and a seed set $S \subseteq V$, we have $\sigma_G(S) \leq |E(S, V \setminus S)| + |S|$, where $E(S, V \setminus S)$ is the set of edges between S and $V \setminus S$.*

The proof of Theorem 3.1 is deferred to Sect. 4.1. We have obtained exactly the same upper bound in Theorem 3.1 for the weighted independent cascade model as well. As for the uniform independent cascade model, we have obtained an upper bound in terms of $E(S, V \setminus S)$, the parameter p and the maximum degree of the vertices in the graph. These results are discussed in Appendix B.

The theorem below shows that the linear threshold INFMAX is APX-hard, even with undirected graphs. We have also obtained the same APX-hardness results for both the uniform independent cascade model and the weighted independent cascade model, which are discussed in Appendix C.

THEOREM 3.2. *There exists a universal constant $\tau \in (0, 1)$ and a function $T : \mathbb{Z}^+ \rightarrow \mathbb{R}^+$ such that, considering the INFMAX problem $(G = (V, E), k)$ with the linear threshold model ULT_G on an undirected graph G , it is NP-hard to distinguish between the following two cases:*

- YES: there exists a seed set S with $|S| = k$ such that $\sigma_G(S) \geq T(n)$;
- NO: for any seed set S with $|S| = k$, we have $\sigma_G(S) \leq (1 - \tau)T(n)$.

PROOF. Let $G_\phi = (V_\phi, E_\phi)$ be the INDSET instance with $|V_\phi| = 3n$ and d be the upper bound of the degrees. We construct the INFMAX instance $G = (V, E)$ as follows. Let D be a sufficiently large integer to be decided later. For each vertex $v \in V_\phi$, create $D + d - \deg(v)$ dummy vertices and connect them to v . We will view G_ϕ as a subgraph of G . By our construction, all vertices in V_ϕ have degree $D + d$ in G . Set $k = n$.

When the INDSET instance G_ϕ is a YES instance, there is an independent set of size n in G_ϕ . If we pick the corresponding n vertices in G as seeds, denoted by S , the expected number of infected vertices after two rounds of the cascade is at least $n + nD + \frac{ndD}{d+D}$. To see this, first notice that, if a vertex $v \in V_\phi$ is infected, then all the dummy vertices connected to it will be infected in the next round, as these vertices have degree 1. Let $\partial S = \{v \in V_\phi \setminus S \mid \exists s \in S : (s, v) \in E_\phi\}$ be the set of all non-dummy vertices that are connected to a seed, and $\partial^d S = \{v \in V \setminus V_\phi \mid \exists s \in S : (s, v) \in E\}$ be the set of all dummy vertices that are connected to a seed. By our construction, we have $|E(S, V_\phi \setminus S)| + |\partial^d S| = n(d + D)$ (since the left hand side is exactly the number of edges between S and $V \setminus S$ and each vertex in the independent set S has degree exactly $d + D$) and $|\partial^d S| \geq nD$ (since each vertex in V_ϕ has at least D dummy neighbors). For each $v \in \partial S$, let δ_v be the number of v 's neighbors in S . In the first round, all vertices in $\partial^d S$ will be infected with probability 1, and each vertex $v \in \partial S$ will be infected with probability $\frac{\delta_v}{d+D}$. If a vertex in ∂S is infected, it will infect all the dummy vertices connected to it in the second round, and there are at least D such dummy vertices. Thus, the total number of vertices that are infected in the first two rounds is at least

$$\begin{aligned} \sum_{v \in \partial^d S} 1 + \sum_{v \in \partial S} \frac{\delta_v}{d+D} \cdot D &= |\partial^d S| + |E(S, V_\phi \setminus S)| \frac{D}{d+D} \\ &= nD + \left(|\partial^d S| - nD\right) + |E(S, V_\phi \setminus S)| \frac{D}{d+D} \\ &\geq nD + \left(|\partial^d S| - nD\right) \frac{D}{d+D} + |E(S, V_\phi \setminus S)| \frac{D}{d+D} \quad (\text{since } |\partial^d S| \geq nD) \\ &= nD + nd \frac{D}{d+D}. \quad (\text{since } |E(S, V_\phi \setminus S)| + |\partial^d S| = n(d+D)) \end{aligned}$$

Therefore, we have

$$\sigma_G(S) \geq n + nD + \frac{ndD}{d+D}. \quad (1)$$

When the INDSET instance G_ϕ is a NO instance, there are at least γn edges between any set of n vertices in G_ϕ . We first note that we can assume without loss of generality that seeds are chosen in $V_\phi \subseteq V$.⁹ Next, consider any seed set $S \subseteq V_\phi \subseteq V$ with $|S| = n$. We have $|E(S, V \setminus S)| \leq n(d+D) - \gamma n$. By Theorem 3.1, this implies

$$\sigma_G(S) \leq n + nD + nd - \gamma n, \quad (2)$$

for any S with $|S| = n$.

Lastly, it is possible to choose D such that

$$1 + D + \frac{dD}{d+D} > 1 + D + d - \gamma,$$

as $\frac{dD}{d+D}$, having limit d when $D \rightarrow \infty$, can be larger than $d - \gamma$ for sufficiently large D . We choose D satisfying this, and notice that D is a universal constant since d and γ are universal constants. Finally, letting

$$\tau = 1 - \frac{1 + D + d - \gamma}{1 + D + \frac{dD}{d+D}} \quad \text{and} \quad T(n) = n + nD + \frac{ndD}{d+D},$$

Eqn. (1) and Eqn. (2) imply the theorem. $\square$

⁹To see this, for (u, v) such that $u \in V_\phi$ and $v \in V \setminus V_\phi$, if u is not seeded, seeding u is strictly more beneficial than seeding v ; if u is seeded, v will be infected with probability 1, while seeding any unseeded vertices in V_ϕ is more beneficial.

We conclude this section by presenting the following APX-hardness result that has *almost perfect completeness*: in the case of a YES instance, the number of infected vertices is $N - o(N)$ deterministically. However, it only holds for general directed graphs. The proof of the theorem makes use of Dinur's PCP theorem as well, but does not depend on the upper bound in Theorem 3.1 as well as any results in Sect. 4. The proof of Theorem 3.3 is available in Appendix C

THEOREM 3.3. *Consider the INFMAX problem $(G = (V, E), k)$ with uniform linear threshold model. There exists a constant $\tau \in (0, 1)$ such that it is NP-hard to distinguish between the following two cases:*

- YES: *there exists a seed set S with $|S| = k$ such that $\Pr(ULT_G(S) \geq N - o(N)) = 1$;*
- NO: *for any seed set S with $|S| = k$, we have $\sigma_G(S) \leq (1 - \tau)N$.*

4 LIFTS, COUPLINGS AND UPPER BOUNDS

In this section, we define the *lift* of an undirected graph G with respect to a vertex set $A \subseteq V$, which is a new undirected graph $\widehat{G}_A$ that has the same vertex set A with G plus a lot of new vertices. We will then define a coupling between sampling live-edges in G and sampling live-edges in $\widehat{G}_S$ (the lift of G with respect to the seeds S). Given the seed set S , this coupling reveals an upper bound of $\sigma_G(S)$ for each of the linear threshold model, the uniform independent cascade model and the weighted independent cascade model on undirected graphs. In particular, we can show that $\sigma_G(S) \leq \sigma_{\widehat{G}_A}(S)$ for all these models. These bounds will imply the APX-hardness results for all these three models, and the upper bounds for the linear threshold model and the weighted independent cascade model are also the key in the INFMAX heuristic presented in Sect. 5.

Let

$$\mathcal{P}_A = \{P = ((v_1, v_2), (v_2, v_3), \dots, (v_{t-1}, v_t)) : v_1 \in A; v_2, \dots, v_t \notin A; \forall i \neq j : v_i \neq v_j\}$$

be the set of all simple paths P that start from vertices in A but never come back to A .

Definition 4.1. Given an undirected graph $G = (V, E)$ and $A \subseteq V$, the *lift* of G with respect to A , denoted by $\widehat{G}_A = (\widehat{V}, \widehat{E})$, is an undirected graph defined as follows.

- The vertex set is $\widehat{V} = A \cup V_P$, where $V_P = \{v_P : P \in \mathcal{P}_A\}$ is the set of vertices corresponding to the simple paths in $\mathcal{P}_A$.
- For each $u \in A$ and $v_P \in V_P$, include $(u, v_P) \in \widehat{E}$ if P is a path of length 1 that starts from u ; for each $v_{P_1}, v_{P_2} \in V_P$, include (v_{P_1}, v_{P_2}) if $|P_2| = |P_1| + 1$ and P_2, P_1 share the first $|P_1|$ common edges (or $|P_1| = |P_2| + 1$ and P_1, P_2 share the first $|P_2|$ common edges, since $\widehat{G}_A$ is undirected); do not include any edge between any two vertices in A .

It is easy to see that the same vertex set A is in both $\widehat{G}_A$ and G , and $\widehat{G}_A$ is a *forest* with roots being exactly those vertices in A . The vertices in the tree rooted at $u \in A$ in $\widehat{G}_A$ correspond to all the paths in $\mathcal{P}_A$ starting at u . For any path $P \in \mathcal{P}_A$ with v being its ending vertex, $\deg(v_P)$ in $\widehat{G}_A$ is less than or equal to $\deg(v)$ in G . Specifically, $\deg(v) - \deg(v_P)$ equals to exactly the number of v 's neighbors that are in A .

Next, we present a coupling argument to show that $\sigma_G(S) \leq \sigma_{\widehat{G}_S}(S)$ for both cascade models. Intuitively, we somehow have separated different “trends” of the cascade on G by considering the corresponding cascade process on G 's tree-like counterpart $\widehat{G}_S$. If a seed s infects vertices $v_1, v_2, \dots, v_t$ one by one along the path

$$P_t := ((s, v_1), (v_1, v_2), \dots, (v_{t-1}, v_t)),$$

it corresponds to the case that the same seed s in the lift $\widehat{G}_S$ infects $v_{P_1}, v_{P_2}, \dots, v_{P_t}$ one by one, where

$$\begin{aligned} P_1 &:= ((s, v_1)), \\ P_2 &:= ((s, v_1), (v_1, v_2)), \\ &\vdots \\ P_t &:= ((s, v_1), (v_1, v_2), \dots, (v_{t-1}, v_t)). \end{aligned}$$

We will define the coupling describing the above correspondence. Moreover, as we will see later, $\widehat{G}_S$ contains many more vertices than G , which potentially produces more infected vertices.

Let $\Psi : E \rightarrow 2^{\widehat{E}}$ be the function mapping an undirected edge in G to its counterparts in $\widehat{G}_A$:

$$\Psi(e) = \begin{cases} \emptyset & \text{if } e = (u, v) \text{ for } u, v \in A \\ \{(u, v_P) \mid P = ((u, v))\} & \text{if } e = (u, v) \text{ for } u \in A, v \notin A \\ \{(v_{P_1}, v_{P_2}) \mid P_2 = (P_1, e)\} & \text{Otherwise.} \end{cases}$$

Notice that in the above definition, $\Psi(e)$ contains only a single edge (u, v_P) with $P = ((u, v))$ being the length-one path connecting u, v if $u \in A$ and $v \notin A$, while $\Psi(e)$ contains the set of all (v_{P_1}, v_{P_2}) such that P_2 is obtained by appending e to P_1 . Let $\Phi : V \rightarrow 2^{\widehat{V}}$ represent the vertex correspondence:

$$\Phi(v) = \begin{cases} \{v\} & \text{if } v \in A \\ \{v_P \mid P \text{ ends at } v\} & \text{Otherwise.} \end{cases}$$

From our definition, it is easy to see that $\Psi(e_1) \cap \Psi(e_2) = \emptyset$ if $e_1 \neq e_2$, and $\Phi(u) \cap \Phi(v) = \emptyset$ if $u \neq v$. Moreover, since $\mathcal{P}_A$ contains only simple paths, for any vertex v and edge e in G , each path in $\widehat{G}_A$ connecting a root in A to a leaf (recall that $\widehat{G}_A$ is a forest with the set of roots exactly A) can intersect each of $\Psi(e)$ and $\Phi(v)$ at most once.¹⁰ We will use this fact multiple times later.

4.1 Upper Bound for the Linear Threshold Model

The following lemma has been shown by Lim et al. [23]. We include a simple proof here for the completeness.

LEMMA 4.2. *Consider a uniform linear threshold model ULT_T with the graph T being an undirected tree. Let $S = \{s\}$ be a seed set containing only a single seed s . We have $\sigma_T(S) = \deg(s) + 1$.*

PROOF. We assume without loss of generality that T is rooted at s . Consider an arbitrary vertex $v \neq s$ at the second last level with children $v_1, \dots, v_t$ being leaves of T . We have $\deg(v) = t + 1$. Suppose v 's parent u is infected by s with probability x ($x = 1$ if $u = s$). Then v will be infected with probability $x/(t + 1)$, and all of $v_1, \dots, v_t$, having degree 1, will be infected with probability 1 if v is infected. Therefore, the expected number of infected vertices in the subtree rooted at v is $\frac{x}{t+1}(1 + t) + (1 - \frac{x}{t+1}) \cdot 0 = x$. This suggests that, if we contract the subtree rooted at v to a single vertex v , the expected total number of infected vertices stays the same for this change of the graph T , since the degree of v becomes 1 after this contraction, making the infection probability of v be

¹⁰To see this for each $\Psi(e)$, suppose for the sake of contradiction that the path from v_P to the root contains two edges $(v_{P_1}, v_{P_2}), (v_{P_3}, v_{P_4})$ such that $(v_{P_1}, v_{P_2}), (v_{P_3}, v_{P_4}) \in \Psi(e)$ for some edge e . Assume without loss of generality that the order of the four vertices on the path according to the distances to the root is $(v_{P_1}, v_{P_2}, v_{P_3}, v_{P_4})$. It is easy to see from our construction that $P_1 \subsetneq P_2 \subsetneq P_3 \subsetneq P_4$. As a result, $(v_{P_1}, v_{P_2}), (v_{P_3}, v_{P_4}) \in \Psi(e)$ implies that P_2 is the path obtained by appending e to P_1 , and P_4 , containing P_2, P_3 , is obtained by appending e to P_3 , which further implies that P_4 is a path that uses the edge e twice, contradicting to our definition that $\mathcal{P}_S$ contains only simple paths.

The corresponding claim for each $\Phi(v)$ can be shown similarly.

the same as that of u , which is x . We can keep doing this contraction until T becomes a star with center s , and the expected number of infected vertices remains the same. The lemma follows. $\square$

With Lemma 4.2 and a coupling argument, we can show Lemma 4.3 below.

LEMMA 4.3. *Given a uniform linear threshold model ULT_G with an undirected graph $G = (V, E)$ and a seed set $S \subseteq V$, we have $\sigma_G(S) \leq \sigma_{\widehat{G}_S}(S)$.*

PROOF. First of all, we modify the graph $\widehat{G}_S$ such that $\deg(v_P) = \deg(v)$ for each $v_P \in \Phi(v)$. We have seen that $\deg(v) - \deg(v_P) \geq 0$, and so we can add $(\deg(v) - \deg(v_P))$ dummy vertices that connect to v_P only. Notice that Lemma 4.2 ensures this modification does not change the expected number of infected vertices: for each $s \in S$, the degree of s in $\widehat{G}_S$ will be the same as the degree of s in G after removing all internal edges of S (the vertices that are adjacent to s in $\widehat{G}_S$ are exactly those paths of length 1 in G , which corresponds to exactly the neighbors of s in G); thus, the tree rooted at s will have exactly $\deg(s) + 1$ infected vertices in expectation, with or without modification (the resultant graph is still a forest after connecting those dummy vertices). We will let $\widehat{G}_S$ be this modified graph from now on.

We will define a coupling between the process of revealing live-edges in G and the process of revealing live-edges in $\widehat{G}_S$. We shall ignore those edges that are internal to S in G , as they do not affect the number of infected vertices. Let χ_G be the edge-revelation process in G , and $\chi_{\widehat{G}_S}$ be the edge-revelation process in $\widehat{G}_S$, where in both processes, each edge is viewed as two anti-parallel directed edges, and we always reveal all the incoming edges for a vertex simultaneously by choosing exactly one incoming edge uniformly at random (this is according to Theorem 2.6 and the remark below Definition 2.7). We will couple χ_G with another edge-revelation process $\chi'_{\widehat{G}_S}$ of $\widehat{G}_S$.

We consider the following coupling. In each iteration where all the incoming edges of v , denoted by $(u_1, v), (u_2, v), \dots, (u_{\deg(v)}, v)$, are revealed such that exactly one of them is live, we reveal all the incoming edges for each $v_P \in \Phi(v)$ as follows.

- For each P' such that $v_{P'}$ is a neighbor of v_P , there must exist $u_i \in \{u_1, \dots, u_{\deg(v)}\}$ such that either that P' is obtained by appending (v, u_i) to P or that P is obtained by appending (u_i, v) to P' . Reveal the directed edge $(v_{P'}, v_P)$ such that it is live if and only if (u_i, v) is live in G .
- If there is a live edge $(v_{P'}, v_P)$ revealed in the above step, make all the remaining directed edges connecting the dummy vertices to v_P not be live. If no live edge is revealed in the above step, choose exactly one of the directed edges connecting the dummy vertices to v_P as the live edge uniformly at random.

This defines a coupling between χ_G and $\chi'_{\widehat{G}_S}$. It is easy to check that each $v_P \in \widehat{V}$ chooses exactly one of its incoming edges uniformly at random in this coupling, which is the same as it is in the process $\chi_{\widehat{G}_S}$. The difference is that, there are dependencies between the revelations of incoming edges for different vertices in $\widehat{G}_S$: if both $v_P, v_{P'} \in \widehat{V}$ belongs to the same $\Phi(v)$ for some $v \in V$, the incoming edges for v_P and $v_{P'}$ are revealed in the same iteration and in the same way.

Although the two processes $\chi'_{\widehat{G}_S}$ and $\chi_{\widehat{G}_S}$ are not the same, we will show that the expected number of vertices that are reachable from S by live edges is the same in both $\chi'_{\widehat{G}_S}$ and $\chi_{\widehat{G}_S}$. It suffices to show that, for each $v_P \in \widehat{V}$, all the vertices in the path connecting v_P to the root of the tree v_P is in are considered independently (meaning that the incoming edges for v_{P_1} on the path are revealed independently to the revelations of the incoming edges of v_{P_2}), since this would imply that the probability v_P is connected to a seed is the same in both $\chi'_{\widehat{G}_S}$ and $\chi_{\widehat{G}_S}$, and the total number of vertices reachable from S by live edges is the same by the linearity of expectation. We only need

to show that there do not exist two vertices on this path that are in the same set $\Phi(v)$ for some $v \in E$, since the incoming edges of each $v_{p_1} \in \Phi(v_1)$ are revealed independently to the revelations of the incoming edges of each $v_{p_2} \in \Phi(v_2)$ whenever $v_1 \neq v_2$. This is true due to that all the paths in $\mathcal{P}_S$ are simple paths, as remarked in the paragraph below where we define function $\Phi(\cdot)$.

Following the same analysis before, we can show that the number of the vertices reachable from S in χ_G is always at most the number of vertices reachable from S in $\chi'_{\widehat{G}_S}$. The lemma concludes here. $\square$

Now we are ready to prove Theorem 3.1.

PROOF OF THEOREM 3.1. For each $s \in S$, let n_s be the number of s 's incident edges with the other ends in $V \setminus S$. It is easy to see that $\widehat{G}_S$ is a forest such that each $s \in S$ has exactly degree n_s . By Lemma 4.3 and Lemma 4.2,

$$\sigma_G(S) \leq \sigma_{\widehat{G}_S}(S) = \sum_{s \in S} (n_s + 1) = |E(S, V \setminus S)| + |S|,$$

which implies the theorem. $\square$

In particular, if we take $S = \{s\}$ in the theorem above, $\deg(s) + 1$ in Lemma 4.2 becomes an upper bound. Notice that the upper bound in the following corollary has also been observed in [23].

COROLLARY 4.4. *Consider a uniform linear threshold model ULT_G with an undirected graph G . Let $S = \{s\}$ be a seed set containing only a single seed s . We have $\sigma_G(S) \leq \deg(s) + 1$.*

This suggests that, in the linear threshold model on undirected graphs, trees provide the most number of infections, and adding more edges to a tree may only reduce the total number of infections.

By “lifting” in a more delicate way, which we called a *boundary lift*, the upper bound in Theorem 3.1 can be further refined to

$$\sigma_G(S) \leq |S| + \sum_{v \in V \setminus S} \frac{\delta_v}{\deg(v)} (1 + \deg(v) - \delta_v), \quad (3)$$

where δ_v is the number of v 's neighbors in S . The refined upper bound, although will not be used in showing the APX-hardness result, will yield a local greedy heuristic with better performance than the heuristic based on upper bound in Theorem 3.1.

Notice that the upper bound above indeed provides a tighter upper bound compared to Theorem 3.1, as

$$|E(S, V \setminus S)| = \sum_{v \in \partial S} \delta_v \geq \sum_{v \in V \setminus S} \frac{\delta_v}{\deg(v)} (1 + \deg(v) - \delta_v).$$

The proof of Eqn. (3) is available in Appendix D. In Appendix D, we will see that the same upper bound also holds for the weighted independent cascade model.

4.2 Upper bounds for the independent cascade model

As we will discuss in Appendix B, the upper bound in Lemma 4.3 holds for both the uniform independent cascade model and the weighted independent cascade model as well. Moreover, Lemma 4.2 and Corollary 4.4 hold for the weighted independent cascade model, with exactly the same proofs.

A similar coupling argument is used to show the same upper bound in Lemma 4.3 for the two independent cascade models. The difference is that we can couple the two edge-revelation processes, χ_G and $\chi'_{\widehat{G}_S}$, by *edges* for the independent cascade model, while the two processes must be coupled

by *vertices* in the linear threshold model as it is in the proof of Lemma 4.3, This is because that, in the edge-revelations for the linear threshold model, the incoming edges of a single vertex are revealed dependently, with one of them being live (Theorem 2.6).

4.3 Discussions about Scalability of Reverse-Reachable-Set-Based Algorithms

As mentioned in Sect. 1.2, reverse-reachable-set-based algorithms give $(1 - 1/e - \epsilon)$ approximation for INFMAX, and are considered as the state-of-the-art in that they produce high quality seeds (almost as good as the greedy algorithm) and are moderately scalable.

In all those reverse-reachable-set-based algorithms, including RIS, TIM⁺, IMM, a sufficient number of reverse reachable sets are sampled. Each reverse reachable set is sampled as follows: first, a vertex v is sampled uniformly at random; second, sample the live edges in the graph according to Theorem 2.2 or Theorem 2.6 (undirected graphs are treated as directed graphs with anti-parallel edges); lastly, the reverse reachable set consists of exactly those vertices from which v is reachable.

After collecting sufficiently many reverse reachable sets, the algorithms choose k seeds that attempt to cover as many reverse reachable sets as possible (we say a reverse reachable set is covered if it contains at least 1 seed), and this is done by a greedy maximum coverage way: iteratively select the seed that maximizes the extra number of reverse reachable sets covered by this seed.

The three algorithms RIS, TIM⁺, IMM differ in the number of reverse reachable sets that is needed in order to preserve the $(1 - 1/e - \epsilon)$ approximation ratio.

Arora et al. [2] observed that these algorithms are much more scalable on the linear threshold model and the weighted independent cascade model than on the uniform independent cascade model. In particular, Arora et al. [2] points out that IMM requires significantly more memory for the uniform independent cascade model compared to the weighted independent cascade model (see Figure 1 in their paper). This indicates that the average size of a random reverse reachable set for the uniform independent cascade model is significantly larger.

Notice that most of the networks used by Arora et al. [2] are undirected. Our upper bounds in Corollary 4.4 and its counterpart for the weighted independent cascade model provide a sound theoretical justification of this phenomenon. For the uniform independent cascade model, p is usually set to 0.1 in the simulations. If a graph contains a lot of well-connected vertices with degrees significantly higher than 10, as it is the case in large social networks, then the size of a random reverse reachable set is likely to be large. On the other hand, Corollary 4.4 and its counterpart for the weighted independent cascade model suggest that, for the linear threshold model and the weighted independent cascade model, the expected size of a reverse reachable set is upper-bounded by the degree of the randomly chosen vertex plus 1, which is small.

5 HIGHLY SCALABLE, EMPIRICALLY GOOD HEURISTICS

For the linear threshold model with undirected graphs, Theorem 3.1 and Eqn. (3) provide upper bounds based only on the local graph structure around the seed set S and the vertices that are adjacent to the seeds. Towards proving both theorems, the (boundary) lift of the graph G with respect to the seed set S has been considered, where the network G has been modified to a forest. We have also seen that tree structures provide the maximum numbers of infections, and adding more edges can only decrease the expected number of infections. Since most of the networks in our real life are sparse, it is very possible that tree structures also provide good *approximations* to the number of infected vertices, and the seeds S maximizing $E(S, V \setminus S)$ or $\sum_{v \in V \setminus S} \frac{\delta_v}{\deg(v)} (1 + \deg(v) - \delta_v)$ as appeared in the two upper bounds are likely to provide good expected number of infections. This motivates our *local greedy heuristics* in this section.

5.1 Local Greedy Heuristics

We consider three local features of a seed set S :

Vertex Cut: The number of vertices that are adjacent to S , denoted by $f_v(S) = |\partial S|$.

Edge Cut: The number of edges between S and $V \setminus S$, denoted by $f_e(S) = |E(S, V \setminus S)|$.

Boundary Expansion: The second term in the upper bound in Eqn. (3), denoted by $f_b(S) = \sum_{v \in V \setminus S} \frac{\delta_v}{\deg(v)} (1 + \deg(v) - \delta_v) = \sum_{v \in \partial S} \frac{\delta_v}{\deg(v)} (1 + \deg(v) - \delta_v)$.

We also consider mixtures of them:

Mixture of Vertex Cut and Edge Cut: $f_{ve}^\lambda(S) = \lambda f_v(S) + (1 - \lambda) f_e(S)$, and

Mixture of Vertex Cut and Boundary Expansion: $f_{vb}^\lambda(S) = \lambda f_v(S) + (1 - \lambda) f_b(S)$,

where λ is a parameter set in the interval $(0, 1)$.

Algorithm 1: Local greedy heuristic with objective function f being one of $\{f_v, f_e, f_b, f_{ve}^\lambda, f_{vb}^\lambda\}$

input : An undirected graph $G = (V, E)$ and an integer k

initialize $S = \emptyset$;

for $i = 1, \dots, k$ **do**

 find $v \in V \setminus S$ maximizing $f(S \cup \{v\}) - f(S)$;

 // follow Table 1

 update $S \leftarrow S \cup \{v\}$;

return S

For a chosen objection function $f \in \{f_v, f_e, f_b, f_{ve}^\lambda, f_{vb}^\lambda\}$, we iteratively choose the seed that maximizes the marginal gain of f , the algorithm is shown in Algorithm 1. Notice that the marginal gain $f(S \cup \{v\}) - f(S)$ can be easily computed for each of the five objective functions as shown in Table 1 (where the marginal gain for f_b in the table follows from straightforward calculations, which are omitted here). In particular, the marginal gains for all these objectives can be computed by only looking at v 's neighbors, and we only need to check each neighbor's degree and the number of each neighbor's neighbors that are already chosen as seeds, which are all local properties that can be stored and updated in a look-up table.

Objective Function	How to Compute Marginal Gain $f(S \cup \{v\}) - f(S)$
f_v	+1 for each v 's neighbor in $V \setminus (S \cup \partial S)$; -1 if $v \in \partial S$
f_e	+1 for each v 's neighbor in $V \setminus S$; -1 for each v 's neighbor in S
f_b	equals to $1 - \frac{\delta_v}{\deg(v)} (1 + \deg(v) - \delta_v) + \sum_{u \in \Gamma(v) \setminus S} \left(1 - 2 \frac{\delta_u}{\deg(u)}\right)$
f_{ve}^λ	equals to $\lambda (f_v(S \cup \{v\}) - f_v(S)) + (1 - \lambda) (f_e(S \cup \{v\}) - f_e(S))$
f_{vb}^λ	equals to $\lambda (f_v(S \cup \{v\}) - f_v(S)) + (1 - \lambda) (f_b(S \cup \{v\}) - f_b(S))$

Table 1. Computing marginal gain for each objective function

5.2 Experiment Setup

We implement the experiments on seven undirected graphs, shown in Table 2. All of our datasets come from [21], and these networks are also popular choices in other empirical work.

All the experiments are implemented on a laptop with an Intel i7 processor and a 16GB memory. We compare the best one of our five local greedy heuristics (with our five different objective functions) to the standard greedy algorithm implemented using 1,000,000 reverse reachable sets as defined in Sect. 4.3. Note that this is more than the number of the reverse reachable sets that

Dataset	Number of Vertices	Number of Edges	Average Degree
CA-GrQc	5,242	14,490	5.53
ego-facebook	1,034	26,749	51.74
Nethept	15,233	31,387	4.12
CA-HepPh	12,008	118,505	19.73
DBLP	317,080	1,049,866	6.62
com-YouTube	1,134,890	2,987,624	5.26
LiveJournal	3,997,962	34,681,189	17.35

Table 2. Datasets for experiments

are generated when standard heuristics (e.g., RIS, TIM⁺ or IMM) are implemented in practice.¹¹ We also compare our heuristics to the simple degree heuristic that chooses the k vertices with highest degrees, and the degree discount heuristics in [6] which iteratively finds the vertex with highest degree and removes it from the graph¹², as they are similar to our heuristics in that only local features of the seeds are ever looked at. This algorithm is called “the single degree discount heuristic” in [6]. Another variant of this heuristic designed specifically for the uniform independent cascade model, named “degree discount IC”, was also presented, and Chen et al. [6] observed that, surprisingly, the seeds output by this variant, although aiming for the uniform independent cascade model, works well for the linear threshold model as well. We will use the words `degreeDiscount` and `degreeDiscountIC` for each heuristic respectively, and we will compare our heuristic to both of them. More details about `degreeDiscountIC` is available in Appendix E.

On those small networks (the first six networks in Table 2), we set the number of seeds $k = 200$ with the graph plotted for every seed (except for Fig. 1 where we are comparing different values of λ for f_{ve}^λ and f_{vb}^λ). For some large networks (the last three networks in Table 2), we set the number of seeds $k = 2000$.

5.3 Results

Figure 1 shows the performance of Algorithm 1 with the objective functions f_{ve}^λ and f_{vb}^λ for $\lambda = 0.9, 0.7, 0.5, 0.3, 0.1$ respectively, on the dataset CA-HepPh with the linear threshold model. We notice that the seeds quality decreases as the value of λ decreases, for both f_{ve}^λ and f_{vb}^λ . This indicates that the local greedy heuristics with more emphasis on vertex cut are better. We observe the same phenomenon for the remaining six datasets. We will only consider $\lambda = 0.9$ for both f_{ve}^λ and f_{vb}^λ from now on.

Although it seems that our observation in Fig. 1 indicates that vertex cut is a better predictor for the seeds quality, setting $\lambda = 1$, i.e., implementing Algorithm 1 with the vertex cut objective f_v , is outperformed by both $f_{ve}^{0.9}$ and $f_{vb}^{0.9}$. Figure 2 compares the performances of $f_v, f_e, f_b, f_{ve}^{0.9}, f_{vb}^{0.9}$ on the dataset CA-HepPh, and the mixture of boundary expansion and vertex cut $f_{vb}^{0.9}$ performs the best and the edge cut f_e performs the worst. These have been observed in the remaining six datasets as well, except for ego-facebook where the vertex cut is worse than the edge cut, while $f_{vb}^{0.9}$ is still the best. Notice that it is not surprising that f_b performs better than f_e and $f_{vb}^{0.9}$ performs

¹¹Our implementation with 1,000,000 reverse reachable sets produce slightly better quality of seed sets than those of TIM⁺ and IMM as reported in [2], where TIM⁺ and IMM are tested on the four networks, Nethept, HepPh, DBLP, Youtube, which we also used in our experiment.

¹²Notice the difference between our edge cut heuristic and the degree discount heuristic: if we start from the sum of the degrees of all the seeds, each internal edge is punished by -1 in the degree discount heuristic, and is punished by -2 in our edge cut heuristic (when adding all the degrees, each edge has been over-counted twice when computing the edge cut).

Fig. 1. Comparing the performances of the local greedy heuristics with objective functions f_{ve}^λ (left) and f_{vb}^λ (right) with $\lambda = 0.9, 0.7, 0.5, 0.3, 0.1$ on dataset CA-HepPh under the linear threshold model.

Fig. 2. Comparing the performances of the local greedy heuristics with the five objective functions $f_e, f_v, f_{ve}^{0.9}, f_b, f_{vb}^{0.9}$ on CA-HepPh under the linear threshold model.

better than $f_{ve}^{0.9}$, as we have remarked after Eqn. (3) that the boundary expansion provides a better upper bound than the edge cut. As a result, we recommend to use $f_{vb}^{0.9}$ in practice.

Finally, we compare, on all the datasets, the performance of $f_{vb}^{0.9}$ to the greedy maximum coverage algorithm on 1,000,000 reverse reachable sets, the two degree discount heuristics `degreeDiscount` and `degreeDiscountIC`, and the algorithm that simply picks the k vertices with highest degrees. Figure 3 shows the results on the seven datasets. In all the experiments, our heuristic has slightly worse but almost the same performance with the greedy maximum coverage algorithm on reverse reachable sets, and our heuristic outperforms the two degree discount heuristics and the heuristic picking k vertices with highest degrees. In particular, our algorithm outperforms the two degree discount heuristics and the max-degree heuristic significantly in some of the datasets. On average, for the linear threshold model, the performance of our heuristic is 97.52% of the performance of the much slower greedy algorithm (on 1,000,000 reverse reachable sets).¹³ However, our heuristic does not require any MCMC simulations or reverse reachable sets samplings, and thus is much more scalable.

Since the upper bounds in Theorem 3.1 and Eqn. (3) also hold for the weighted independent cascade model, we have tried the local greedy heuristics on this model as well on the same datasets with the same experiment setup. On average, the performance of the local greedy heuristic with $f_{vb}^{0.9}$ is 97.39% of the performance of the greedy maximum coverage algorithm on 1,000,000 reverse reachable sets, and the details are in Appendix E.

¹³This percentage is averaging over the experiments with exactly 200 seeds. If including those with 2000 seeds, the percentage is 97.20%.

Fig. 3. Comparing the local greedy heuristic with $f_{vb}^{0.9}$ to the greedy maximum coverage algorithm on 1,000,000 reverse reachable sets (labelled as “greedy (RR sets)”), degreeDiscount, degreeDiscountIC, and the algorithm that simply pick k vertices with highest degrees (labelled as “maxDegree”). The cascade model is the linear threshold model.

On the other hand, the local greedy heuristic is less promising for the uniform independent cascade model (only 86.26% of the greedy algorithm). This is not surprising, as our local characterization does not applied to this model (for example, the upper bound in Eqn. (3) does not hold). The experiment results for the uniform independent cascade model is also available in Appendix E.

6 CONCLUSION

We have seen that INFMAX is APX-hard for both the linear threshold model and the independent cascade model on undirected graphs. For the linear threshold model, there is still a gap between the upper bound $(1 - 1/e)$ and the lower bound $(1 - \tau)$, while the gap is slightly smaller for the independent cascade model with upper bound $(1 - 1/e + c)$ and lower bound $(1 - \tau)$. A natural open problem is to close these gap by either designing an approximation algorithm, taking advantage of the undirected nature of the graph, that achieves approximation guarantee better than $(1 - 1/e)$ (or $(1 - 1/e + c)$ for the independent cascade model), or to show that the problem is hard to approximate with a larger gap. Another interesting future direction is to study the same problem for the linear threshold model with directed graphs. Notice that the same gap still exists even for this setting.

REFERENCES

- [1] Rico Angell and Grant Schoenebeck. 2017. Don’t be greedy: Leveraging community structure to find high quality seed sets for influence maximization. In *International Conference on Web and Internet Economics*. Springer, 16–29.
- [2] Akhil Arora, Sainyam Galhotra, and Sayan Ranu. 2017. Debunking the myths of influence maximization: An in-depth benchmarking study. In *Proceedings of the 2017 ACM International Conference on Management of Data*. ACM, 651–666.
- [3] Christian Borgs, Michael Brautbar, Jennifer Chayes, and Brendan Lucier. 2014. Maximizing social influence in nearly optimal time. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 946–957.

- [4] Ning Chen. 2009. On the approximability of influence in social networks. *SIAM Journal on Discrete Mathematics* 23, 3 (2009), 1400–1415.
- [5] Wei Chen. 2018. An issue in the martingale analysis of the influence maximization algorithm IMM. In *International Conference on Computational Social Networks*. Springer, 286–297.
- [6] Wei Chen, Yajun Wang, and Siyu Yang. 2009. Efficient influence maximization in social networks. In *ACM SIGKDD*. ACM, 199–208.
- [7] Wei Chen, Yifei Yuan, and Li Zhang. 2010. Scalable Influence Maximization in Social Networks under the Linear Threshold Model. In *2010 IEEE International Conference on Data Mining*. IEEE, 88–97.
- [8] Wei Chen, Yifei Yuan, and Li Zhang. 2010. Scalable influence maximization in social networks under the linear threshold model. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 88–97.
- [9] Suqi Cheng, Huawei Shen, Junming Huang, Guoqing Zhang, and Xueqi Cheng. 2013. Staticgreedy: solving the scalability-accuracy dilemma in influence maximization. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 509–518.
- [10] Irit Dinur. 2007. The PCP theorem by gap amplification. *Journal of the ACM (JACM)* 54, 3 (2007), 12.
- [11] P. Domingos and M. Richardson. 2001. Mining the network value of customers. In *ACM SIGKDD*.
- [12] Uriel Feige. 1998. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM (JACM)* 45, 4 (1998), 634–652.
- [13] Sainyam Ghalotra, Akhil Arora, and Shourya Roy. 2016. Holistic influence maximization: Combining scalability and efficiency with opinion-aware models. In *Conference on Management of Data*. ACM, 743–758.
- [14] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. 2011. Celf++: optimizing the greedy algorithm for influence maximization in social networks. In *Proceedings of the 20th international conference WWW*. ACM, 47–48.
- [15] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. 2011. Simpath: An efficient algorithm for influence maximization under the linear threshold model. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 211–220.
- [16] Kyomin Jung, Wooram Heo, and Wei Chen. 2012. Irie: Scalable and robust influence maximization in social networks. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 918–923.
- [17] David Kempe, Jon M. Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *ACM SIGKDD*. 137–146.
- [18] David Kempe, Jon M. Kleinberg, and Éva Tardos. 2005. Influential Nodes in a Diffusion Model for Social Networks. In *ICALP*. 1127–1138.
- [19] Sanjeev Khanna and Brendan Lucier. 2014. Influence maximization in undirected networks. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 1482–1496.
- [20] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 420–429.
- [21] Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. <http://snap.stanford.edu/data>.
- [22] Qiang Li, Wei Chen, Xiaoming Sun, and Jialin Zhang. 2017. Influence Maximization with ϵ -Almost Submodular Threshold Functions. In *NIPS*. 3804–3814.
- [23] Yongwhan Lim, Asuman Ozdaglar, and Alexander Teytelboym. 2015. A simple model of cascades in networks.
- [24] Elchanan Mossel and Sébastien Roch. 2010. Submodularity of Influence in Social Networks: From Local to Global. *SIAM J. Comput.* 39, 6 (2010), 2176–2188.
- [25] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming* 14, 1 (1978), 265–294.
- [26] Naoto Ohsaka, Takuya Akiba, Yuichi Yoshida, and Ken-ichi Kawarabayashi. 2014. Fast and Accurate Influence Maximization on Large Networks with Pruned Monte-Carlo Simulations.. In *AAAI*. 138–144.
- [27] M. Richardson and P. Domingos. 2002. Mining knowledge-sharing sites for viral marketing. In *ACM SIGKDD*. 61–70.
- [28] Grant Schoenebeck and Biaoshuai Tao. 2017. Beyond Worst-Case (In)approximability of Nonsubmodular Influence Maximization. In *International Conference on Web and Internet Economics*. Springer, 368–382.
- [29] Grant Schoenebeck and Biaoshuai Tao. 2019. Beyond worst-case (in)approximability of nonsubmodular influence maximization. *ACM Transactions on Computation Theory (TOCT)* 11, 3 (2019), 12.
- [30] Youze Tang, Yanchen Shi, and Xiaokui Xiao. 2015. Influence maximization in near-linear time: A martingale approach. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM, 1539–1554.
- [31] Youze Tang, Xiaokui Xiao, and Yanchen Shi. 2014. Influence maximization: Near-optimal time complexity meets practical efficiency. In *SIGMOD international conference on Management of data*. ACM, 75–86.
- [32] Yaron Singer Thibaut Horel. 2016. Maximization of Approximately Submodular Functions. In *NIPS*.

A DINUR'S PCP THEOREM AND ITS IMPLICATIONS

We will use the following version of the PCP theorem which is due to Dinur [10].

THEOREM A.1. *There exist a universal constant integer d and a universal constant $\gamma \in (0, 1)$ such that, given a 3SAT instance ϕ where each variable appears in at most d clauses, it is NP-hard to distinguish between the following two cases:*

- YES: ϕ has a satisfiable truth-assignment;
- NO: for any truth-assignment, at most $(1 - \gamma)$ fraction of clauses can be satisfied.

Dinur's PCP theorem straightforwardly implies Theorem 2.8.

PROOF OF THEOREM 2.8. Given a 3SAT instance ϕ with n clauses such that each variable appears in at most d' clauses, we construct the INDSET instance $G_\phi = (V_\phi, E_\phi)$ as follows: G_ϕ contains $3n$ vertices $\{v_{ij} \mid i = 1, \dots, n; j = 1, 2, 3\}$ where v_{ij} corresponds to the j -th literal in the i -th clause of ϕ ; $(v_{ij}, v_{i'j'}) \in E_\phi$ if $i = i'$ or $v_{ij}, v_{i'j'}$ correspond to two literals such that one is the negation of the other. It is easy to see that if each variable in ϕ appears in at most d' clauses, then each vertex in G_ϕ has degree at most $d = d' + 1$. If the 3SAT instance is a YES instance, then ϕ is satisfiable and has an assignment which satisfies all the n clauses. This assignment, by including one true literal in each clause, yields an independent set of size n in G . If the 3SAT instance is a NO instance, Dinur's PCP theorem implies that there exists a constant $\gamma \in (0, 1)$ such that no assignment satisfies even $(1 - \gamma)n$ clauses. This means that no independent set of size more than $(1 - \gamma)n$ exists, because by including one true literal in each clause for a given assignment yields an independent set with the same number of vertices as satisfying clauses. As a result, any subgraph of size n contains at least γn edges: if the number of edges in certain subgraph with size n is less than γn , removing one endpoint for each edge gives an independent set of size more than $n - \gamma n$, which is a contradiction. $\square$

Dinur's PCP theorem also implies the APX-hardness of the vertex cover problem (VERTEXCOVER) on graphs with constant degree-bound. We will need it for the proof of Theorem 3.3.

THEOREM A.2 (INAPPROXIMABILITY OF VERTEXCOVER). *There exist a universal constant integer d and a universal constant $\gamma \in (0, 1)$ such that, given an integer k and an undirected graph $G = (V, E)$ where the degree of each vertex is bounded by d , it is NP-hard to distinguish between the following two cases:*

- YES: G has a vertex cover of size k ;
- NO: all vertex covers of G have sizes at least $(1 + \gamma)k$.

PROOF. Fix $\gamma > 0$ from Theorem A.1. We will show that Theorem A.2 is true for substituting γ with $\gamma/2$. Given a 3SAT instance ϕ with n clauses, we construct the VERTEXCOVER instance (G_ϕ, k) with $k = 2n$ and G_ϕ being the same graph as it is in the proof of Theorem 2.8. By Theorem 2.8 (or following the same arguments in the proof of Theorem 2.8, if ϕ is a YES instance, G_ϕ has an independent set of size n , picking the remaining $2n$ vertices that are not in this independent set gives a vertex cover of size $k = 2n$. If ϕ is a NO instance, Dinur's PCP theorem implies that no assignment satisfies even $(1 - \gamma)n$ clauses. This means that no independent set of size $(1 - \gamma)n$ exists (because by including one true literal in each clause for a given assignment yields an independent set with the same number of vertices as satisfying clauses). Thus, all vertex covers of G have sizes at least $3n - (1 - \gamma)n = (2 + \gamma)n$. This implies the YES and NO instances differ by a factor of $1 + \gamma/2$. $\square$

B UPPER BOUNDS FOR THE INDEPENDENT CASCADE MODEL

In this section, we derive the upper bounds for the uniform independent cascade model and the weighted independent cascade model by using the same technique in Sect. 4. Given an undirected graph $G = (V, E)$ and $A \subseteq V$, let $\widehat{G}_A$ be the lift of G with respect to A as defined in Definition 4.1.

B.1 Upper Bound for the Uniform Independent Cascade Model

We consider the uniform independent cascade model with p being fixed for the weight of all edges in this subsection.

LEMMA B.1. *Given a uniform independent cascade model UIC_G^p with an undirected graph $G = (V, E)$ and a seed set $S \subseteq V$, we have $\sigma_G(S) \leq \sigma_{\widehat{G}_S}(S)$.*

PROOF. We will define a coupling between the process of revealing live-edges in G and the process of revealing live-edges in $\widehat{G}_S$. We shall ignore those edges that are internal to S in G , as they do not affect the number of infected vertices. Let χ_G be the edge-revelation process in G , and $\chi_{\widehat{G}_S}$ be the edge-revelation process in $\widehat{G}_S$, where each edge is revealed with probability p independently in both processes. We will couple χ_G with another edge-revelation process $\chi'_{\widehat{G}_S}$ of $\widehat{G}_S$.

We consider the following coupling. For each undirected edge e in G , we reveal both of its corresponding two anti-parallel directed edges, and meanwhile reveal the two anti-parallel edges for each undirected edge in $\Psi(e)$ in the following way:

- if $e = (u, v)$ for $u \in S$ and $v \notin S$, then for each undirected edge $(u, v_P) \in \Psi(e)$ where $P = ((u, v))$, make (u, v_P) live if and only if (u, v) is live in χ_G , and make (v_P, u) live if and only if (v, u) is live in χ_G ;
- if $e = (u, v)$ for $u, v \notin S$, then
 - for each undirected edge $(v_{P_1}, v_{P_2}) \in \Psi(e)$ where P_1 ends at u and $P_2 = (P_1, e)$, make (v_{P_1}, v_{P_2}) live if and only if (u, v) is live in χ_G , and make (v_{P_2}, v_{P_1}) live if and only if (v, u) is live in χ_G ;
 - for each undirected edge $(v_{P_1}, v_{P_2}) \in \Psi(e)$ where P_1 ends at v and $P_2 = (P_1, e)$, make (v_{P_1}, v_{P_2}) live if and only if (v, u) is live in χ_G , and make (v_{P_2}, v_{P_1}) live if and only if (u, v) is live in χ_G .

This defines a coupling between χ_G and $\chi'_{\widehat{G}_S}$. Notice that the two processes $\chi'_{\widehat{G}_S}$ and $\chi_{\widehat{G}_S}$ are not the same, because the edges in $\chi'_{\widehat{G}_S}$ are not revealed independently. However, we will show that the expected number of vertices that are reachable from S by live edges is the same in both $\chi'_{\widehat{G}_S}$ and $\chi_{\widehat{G}_S}$, which implies that the expected number of vertices that are reachable from S by live edges in $\chi'_{\widehat{G}_S}$ is still exactly $\sigma_{\widehat{G}_S}(S)$ (since Theorem 2.2 shows that the expected number of vertices that are reachable from S in $\chi_{\widehat{G}_S}$ is $\sigma_{\widehat{G}_S}(S)$).

To see this, it suffices to show that, for each $v_P \in \widehat{V}$, all the directed edges in the path connecting from the root of the tree v_P is in to the vertex v_P are sampled independently, since this would imply that the probability v_P is connected to a seed is the same in both $\chi'_{\widehat{G}_S}$ and $\chi_{\widehat{G}_S}$, and the total number of vertices reachable from S by live edges is the same by the linearity of expectation. We only need to show that there do not exist two edges on this path that are in the same set $\Psi(e)$ for some $e \in E$, since edges in $\Psi(e_1)$ are revealed independently to the revelations of edges in $\Psi(e_2)$ whenever $e_1 \neq e_2$ (this is because e_1, e_2 are revealed independently in χ_G). This is true since paths in $\mathcal{P}_S$ are all simple paths, as remarked earlier.

To conclude the lemma, we will show that the number of the vertices reachable from S in χ_G is always at most the number of vertices reachable from S in $\chi'_{\widehat{G}_S}$. It is easy to see that, if $v \in V$ is connected to S by a path P consisting of live edges, the vertex $v_P \in \widehat{V}$ is also connected to S in $\widehat{G}_S$ by live edges. Thus, there exists at least one vertex in $\Phi(u)$ connected to S by live edges in $\widehat{G}_S$ for each vertex $u \in V$ connected to S by live edges in G . The lemma follows from that $\Phi(u)$'s are non-overlapping. $\square$

Now we are ready to prove the following theorem, which, as a counterpart to Theorem 3.1, provides an upper bound for $\sigma_G(S)$ for the uniform independent cascade model.

THEOREM B.2. *Given a uniform independent cascade model UIC_G^p with a seed set $S \subseteq V$, and assuming $p < \frac{1}{d}$ where d is the maximum degree over all vertices, we have $\sigma_G(S) \leq |S| + \frac{|E(S, V \setminus S)|p}{1-pd}$.*

PROOF. For each seed $s \in S \subseteq V$, let δ_s be the number of s 's neighbors in $V \setminus S$. Let T be a forest such that each vertex $s \in S$ is the root with δ_s children, and each child is a root of an infinite full d -ary subtree. Clearly, $\widehat{G}_S$ is a subgraph of T , so $\sigma_T(S) \geq \sigma_{\widehat{G}_S}(S) \geq \sigma_G(S)$ by Lemma B.1. For each tree in T rooted at $s \in S$, the probability that a vertex on level i is infected is p^i , and there are $\delta_s d^{i-1}$ vertices on level i . Therefore, we have

$$\sigma_T(S) = |S| + \sum_{s \in S} \sum_{i=1}^{\infty} \delta_s d^{i-1} p^i = |S| + \sum_{i=1}^{\infty} |E(S, V \setminus S)| d^{i-1} p^i = |S| + \frac{|E(S, V \setminus S)|p}{1-pd},$$

which implies the theorem. $\square$

Lemma B.1 and Theorem B.2 can be generalized to directed graphs as well. The lift $\widehat{G}_A$ should then be defined as the directed forest such that the root for each tree is the source and the leaves are the sinks. With the same proofs, Lemma B.1 and Theorem B.2 follow for directed graphs, where d in Theorem B.2 becomes the maximum *out-degree*. Since we only need Theorem B.2 for undirected graphs and the generalization to directed graphs is straightforward, we omit the details for directed graphs.

B.2 Upper Bound for the Weighted Independent Cascade Model

Firstly, Lemma 4.2 holds for the weighted independent cascade model as well, with exactly the same proof.

LEMMA B.3. *Consider a weighted independent cascade model WIC_T with the graph T being an undirected tree. Let $S = \{s\}$ be a seed set containing only a single seed s . We have $\sigma_T(S) = \deg(s) + 1$.*

Next, we can show the following lemma by combining ideas in the proofs of both Lemma 4.3 and Lemma B.1.

LEMMA B.4. *Given a weighted independent cascade model WIC_G with an undirected graph $G = (V, E)$ and a seed set $S \subseteq V$, we have $\sigma_G(S) \leq \sigma_{\widehat{G}_S}(S)$.*

PROOF. We first modify the graph $\widehat{G}_S$ in the same way as described in the first paragraph of the proof of Lemma 4.3. Then, the remaining part of the coupling follows the one in the proof of Lemma B.1. Notice that the modification of $\widehat{G}_S$ ensures that $v \in V$ has the same degree as the degree of each vertex in $\Phi(v)$. This validates the coupling between χ_G and $\chi'_{\widehat{G}_S}$. $\square$

Applying the same arguments in the proof of Theorem 3.1, Lemma B.4 implies Theorem B.5 below.

THEOREM B.5. *Given a weighted independent cascade model WIC_G with an undirected graph $G = (V, E)$ and a seed set $S \subseteq V$, we have $\sigma_G(S) \leq |E(S, V \setminus S)| + |S|$, where $E(S, V \setminus S)$ is the set of edges between S and $V \setminus S$.*

The following corollary of Theorem B.5, as a counterpart to Corollary 4.4, shows that trees also provide the most number of infections for the weighted independent cascade model.

COROLLARY B.6. *Consider a weighted independent cascade model WIC_G with an undirected graph G . Let $S = \{s\}$ be a seed set containing only a single seed s . We have $\sigma_G(S) \leq \deg(s) + 1$.*

Notice that, although Theorem B.2 can be adapted for directed graphs with d in the theorem being the maximum *out-degree*, Theorem 3.1 and B.5 only hold for undirected graphs. The linear threshold model and the weighted independent cascade model on directed graphs have a fundamentally different nature. The local characterization no longer applies and the cascade process is no longer diminishing. As a simple example, if a seed s is put at one end of a directed line, $G = (V, E)$ where $E = \{(s, v_1), (v_1, v_2), \dots, (v_{N-1}, v_N)\}$, then, for both the linear threshold model and the weighted independent cascade model, $\sigma_G(\{s\}) = N + 1$ as all the vertices in G will be infected with probability 1, even if the degree of s is only 1.

C OMITTED APX-HARDNESS RESULTS FOR INFMAX

The theorem below shows that INFMAX on undirected graphs with the independent cascade model is APX-hard, even if all edges have the same weight (i.e., the uniform independent cascade model).

THEOREM C.1. *There exist universal constants $\tau, p \in (0, 1)$ and a function $T : \mathbb{Z}^+ \rightarrow \mathbb{R}^+$ such that, considering the INFMAX problem $(G = (V, E), k)$ with a uniform independent cascade model UIC_G^p on an undirected graph G , it is NP-hard to distinguish between the following two cases:*

- YES: *there exists a seed set S with $|S| = k$ such that $\sigma_G(S) \geq T(|V|)$;*
- NO: *for any seed set S with $|S| = k$, we have $\sigma_G(S) \leq (1 - \tau)T(|V|)$.*

PROOF. The reduction is as follows. Given an INDSET instance $G_\phi = (V_\phi, E_\phi)$ with $|V_\phi| = 3n$ and degree bound d (and a potential independent set of size n is regarded), the undirected graph $G = (V, E)$ is obtained by modifying $G_\phi = (V_\phi, E_\phi)$ as follows: let d be the maximum degree over vertices in G_ϕ , for each $v \in V_\phi$ with $\deg(v) < d$, add $d - \deg(v)$ dummy vertices and make them connect to v only. This makes all the original $3n$ vertices in V_ϕ have the same degree d . The number of seeds k is set to n . We will decide the values for p and τ , and the function $T(\cdot)$ later. In the remaining part of this proof, we will view G_ϕ as a subgraph of G so that $V \setminus V_\phi$ are exactly those dummy vertices.

When the INDSET instance G_ϕ is a YES instance, there is an independent set of size n in G_ϕ . If we pick the corresponding n vertices in G as seeds, denoted by S , the expected number of infected vertices after the first round of the cascade is at least $ndp(1-p)^{d-1}$. This is because the total number of edges between S and $V \setminus S$ is exactly nd (since S is an independent set and each $s \in S$ has degree d), and the expected number of infected vertices for the first round is

$$\begin{aligned}
 \sum_{v \in \partial S} \Pr(v \text{ is infected}) &= \sum_{v \in \partial S} \left(1 - (1-p)^{\delta_v}\right) \\
 &\text{(where } \partial S = \{v \in V \setminus S \mid \exists s \in S : (s, v) \in E\} \text{ and } \delta_v \text{ is the number of } v\text{'s neighbors in } S) \\
 &= \sum_{v \in \partial S} \sum_{i=0}^{\delta_v-1} p(1-p)^i \\
 &\geq \sum_{v \in \partial S} \sum_{i=0}^{\delta_v-1} p(1-p)^{d-1} \quad (\text{since } i \leq \delta_v - 1 \leq d - 1) \\
 &= ndp(1-p)^{d-1},
 \end{aligned}$$

where the last equality follows since $\sum_{v \in \partial S} \sum_{i=0}^{\delta_v-1} 1$ is exactly the number of edges between S and $V \setminus S$, which is nd . Therefore, by only consider the first round of the cascade, we have

$$\sigma_G(S) \geq n + ndp(1-p)^{d-1}, \quad (4)$$

for certain $S \subseteq V$ if G_ϕ is a YES instance.

When the INDSET instance G_ϕ is a NO instance, there are at least γn edges between any set of n vertices in G_ϕ . We first note that we can assume without loss of generality that seeds are chosen in $V_\phi \subseteq V$, as seeding any $w \in V \setminus V_\phi$ is always less beneficial.¹⁴ Next, consider any seed set $S \subseteq V_\phi \subseteq V$ with $|S| = n$. We have $|E(S, V \setminus S)| \leq nd - \gamma n = n(d - \gamma)$. By Theorem B.2, if we set p such that $p < \frac{1}{d}$, this implies

$$\sigma_G(S) \leq n + np \frac{d - \gamma}{1 - dp}, \quad (5)$$

for any S with $|S| = n$.

Next, the inequality

$$d(1 - p)^{d-1} > \frac{d - \gamma}{1 - dp}$$

holds for sufficiently small p , as the left-hand side has limit d and the right-hand side has limit $d - \gamma < d$ when $p \rightarrow 0$. We choose the value of p with $p > 0$ such that this inequality still holds. Notice that p is a universal constant, since d and γ in the equality are universal constants. Let

$$\tau = 1 - \frac{1 + p \frac{d - \gamma}{1 - dp}}{1 + dp(1 - p)^{d-1}}.$$

Our choice of p makes sure $\tau > 0$, and τ is a constant since p , d and γ are all constants. Finally, choosing $T(n) = n + ndp(1 - p)^{d-1}$, it is straightforward to check that Eqn. (4) and Eqn. (5) imply the theorem. $\square$

The APX-hardness for the weighted independent cascade model can be proved in a way similar to proving Theorem 3.2.

THEOREM C.2. *There exists a universal constant $\tau \in (0, 1)$ and a function $T : \mathbb{Z}^+ \rightarrow \mathbb{R}^+$ such that, considering the INFMAX problem $(G = (V, E), k)$ with the weighted independent cascade model WIC_G on an undirected graph G , it is NP-hard to distinguish between the following two cases:*

- YES: *there exists a seed set S with $|S| = k$ such that $\sigma_G(S) \geq T(n)$;*
- NO: *for any seed set S with $|S| = k$, we have $\sigma_G(S) \leq (1 - \tau)T(n)$.*

PROOF. Let $G_\phi = (V_\phi, E_\phi)$ be the INDSET instance, with d being the upper bound of the degrees and $|V_\phi| = 3n$. We construct the INFMAX instance $G = (V, E)$ in exactly the same way as it is in the proof of Theorem 3.2, and the corresponding constant D is decided later.

When the INDSET instance G_ϕ is a YES instance, there is an independent set of size n in G_ϕ . Again, we consider picking the corresponding n vertices in G as seeds, denoted by S , and find a lower bound for $\sigma_G(S)$. Let ∂S and $\partial^d S$ be the same as they are in the proof of Theorem 3.2. For each $v \in \partial S$, let δ_v be the number of v 's neighbors in S again. In the first round, all vertices in $\partial^d S$ will be infected with probability 1, and each vertex $v \in \partial S$ will be infected with probability $1 - (1 - \frac{1}{d+D})^{\delta_v}$. If a vertex in ∂S is infected, it will infect all the dummy vertices connected to it in

¹⁴To see this, for (u, w) such that $u \in V_\phi$ and $w \in V \setminus V_\phi$, if u is not seeded, seeding u is strictly more beneficial than seeding w ; if u is seeded, seeding w adds only $1 - p$ to $\sigma_G(S)$, while seeding any unseeded vertices in V_ϕ is more beneficial.

the second round, and there are at least D such dummy vertices. Putting together, we have

$$\begin{aligned}
\sigma_G(S) &\geq n + \sum_{v \in \partial^d S} 1 + \sum_{v \in \partial S} \left(1 - \left(1 - \frac{1}{d+D}\right)^{\delta_v}\right) D \\
&= n + |\partial^d S| + \sum_{v \in \partial S} \sum_{i=0}^{\delta_v-1} \frac{1}{d+D} \left(1 - \frac{1}{d+D}\right)^i D \\
&\geq n + |\partial^d S| + \sum_{v \in \partial S} \sum_{i=0}^{\delta_v-1} \frac{D}{d+D} \left(1 - \frac{1}{d+D}\right)^d \quad (\text{since } i < \delta_v \leq d) \\
&= n + |\partial^d S| + |E(S, V_\phi \setminus S)| \frac{D}{d+D} \left(1 - \frac{1}{d+D}\right)^d \quad (\text{since } \sum_{v \in \partial S} \sum_{i=0}^{\delta_v-1} 1 = |E(S, V_\phi \setminus S)|) \\
&\geq n + nD + \left(|\partial^d S| - nD + |E(S, V_\phi \setminus S)|\right) \frac{D}{d+D} \left(1 - \frac{1}{d+D}\right)^d \quad (\text{since } |\partial^d S| \geq nD) \\
&= n + nD + \frac{ndD}{d+D} \left(1 - \frac{1}{d+D}\right)^d. \quad (\text{since } |E(S, V_\phi \setminus S)| + |\partial^d S| = n(d+D))
\end{aligned}$$

When the INDSET instance G_ϕ is a NO instance, the same analysis in the proof of Theorem 3.2, coupled with Theorem B.5, implies that

$$\sigma_G(S) \leq n + nD + nd - n\gamma,$$

for any S with $|S| = n$.

Lastly, it is possible to choose D such that

$$1 + D + \frac{dD}{d+D} \left(1 - \frac{1}{d+D}\right)^d > 1 + D + d - \gamma,$$

as $\frac{dD}{d+D} \left(1 - \frac{1}{d+D}\right)^d$, having limit d when $D \rightarrow \infty$, can be larger than $d - \gamma$ for sufficiently large D . We choose D satisfying this, and notice that D is a universal constant since d and γ are universal constants. Finally, letting

$$\tau = 1 - \frac{1 + D + d - \gamma}{1 + D + \frac{dD}{d+D} \left(1 - \frac{1}{d+D}\right)^d} \quad \text{and} \quad T(n) = n + nD + \frac{ndD}{d+D} \left(1 - \frac{1}{d+D}\right)^d,$$

the theorem is implied. $\square$

Finally, we prove Theorem 3.3 using another corollary of Dinur's PCP theorem, namely, Theorem A.2.

PROOF OF THEOREM 3.3. We consider a reduction from VERTEXCOVER. Given a VERTEXCOVER instance $(\bar{G} = (\bar{V}, \bar{E}), \bar{k})$, we will construct an INFMAX instance $(G = (V, E, w), k)$. Let $\bar{n} = |\bar{V}|$, $\bar{m} = |\bar{E}|$, and $W = 2\bar{m}^5\bar{n}^5$ (W can be any sufficiently large even number in this reduction, as long as it is bounded by a polynomial of $\bar{m}$ and $\bar{n}$). We consider exclusively those $\bar{G}$ such that each $v \in \bar{V}$ has degree at most d . Let γ and the notion of YES/NO instance be defined as they are in Theorem A.2. We further assume that $\bar{k} \geq \bar{m}/d$. This can be assume without loss of generality: the total number of edges that can possibly be covered by $\bar{k}$ vertices is at most $\bar{k}d$, so a set of less than $\bar{m}/d$ vertices cannot be a vertex cover and the VERTEXCOVER instance is a trivial NO instance if $\bar{k} < \bar{m}/d$.

Fig. 4. The edge gadget

The number of seeds k is set to $k = \bar{m} + \bar{k}$. The graph G contains $N = \bar{n} + \bar{m}(2 + W)$ vertices and is defined as follows.

- For each $\bar{v} \in \bar{V}$, construct a vertex $v \in V$; for each $(\bar{u}, \bar{v}) \in \bar{E}$, construct $2 + W$ vertices: $e_{u,v}^u, e_{u,v}^v, w_{u,v}^1, w_{u,v}^2, \dots, w_{u,v}^W$.
- For each $(\bar{u}, \bar{v}) \in \bar{E}$, construct $4 + W$ edges:
 - $(u, e_{u,v}^v), (v, e_{u,v}^u), (e_{u,v}^u, e_{u,v}^v), (e_{u,v}^v, e_{u,v}^u)$, each of which has weight 0.5, and
 - $(e_{u,v}^u, w_{u,v}^1), \dots, (e_{u,v}^u, w_{u,v}^{\frac{W}{2}}), (e_{u,v}^v, w_{u,v}^{\frac{W}{2}+1}), \dots, (e_{u,v}^v, w_{u,v}^W)$, each of which has weight 1.

For each $(\bar{u}, \bar{v}) \in \bar{E}$, we have constructed a *gadget* shown in Fig. 4, for which we will denote by $G_{u,v}$, that contains $4 + W$ vertices $u, v, e_{u,v}^u, e_{u,v}^v, w_{u,v}^1, w_{u,v}^2, \dots, w_{u,v}^W$ and $4 + W$ edges defined above. Notice that any two gadgets can only share at most a single vertex v which corresponds to a certain $\bar{v} \in \bar{V}$. More importantly, the infected vertices in $G_{u,v} \setminus \{v\}$ will never infect v as v is connected to the rest of $G_{u,v}$ by a directed edge. This ensures that the infection of any one or more of $e_{u,v}^u, e_{u,v}^v, w_{u,v}^1, w_{u,v}^2, \dots, w_{u,v}^W$ in a gadget $G_{u,v}$ has no impact on the other gadgets. Thus, after a seed set is fixed, we can analyze the cascade in each gadget independently. Finally, it is easy to see that the construction has a polynomial size, so this is a polynomial time reduction.

If the VERTEXCOVER instance is a YES instance, we assume without loss of generality that $\{\bar{v}_1, \dots, \bar{v}_{\bar{k}}\}$ is a vertex cover of $\bar{G}$. Consider the seed set S of size $k = \bar{m} + \bar{k}$ defined as follows:

- S includes $v_1, \dots, v_{\bar{k}}$;
- for each gadget $G_{u,v}$, S includes exactly one of $e_{u,v}^u, e_{u,v}^v$: S includes $e_{u,v}^u$ if $u \in \{v_1, \dots, v_{\bar{k}}\}$, S includes $e_{u,v}^v$ if $v \in \{v_1, \dots, v_{\bar{k}}\}$, and pick an arbitrary one of $e_{u,v}^u, e_{u,v}^v$ to be included in S if $u, v \in \{v_1, \dots, v_{\bar{k}}\}$. Notice that we have either $u \in \{v_1, \dots, v_{\bar{k}}\}$ or $v \in \{v_1, \dots, v_{\bar{k}}\}$ since $\{\bar{v}_1, \dots, \bar{v}_{\bar{k}}\}$ is a vertex cover of $\bar{G}$.

It is easy to see that each gadget $G_{u,v}$ contains at least $3 + W \geq W$ infected vertices after the cascade: assume without loss of generality that $u, e_{u,v}^u \in S$; then $e_{u,v}^v$, having all the two in-neighbors $u, e_{u,v}^u$ infected, will be infected with probability 1; finally, the infection of both $e_{u,v}^u$ and $e_{u,v}^v$ will infect all of $w_{u,v}^1, \dots, w_{u,v}^W$ with probability 1. Therefore, with probability 1, the total number of infected vertices is at least $\bar{m}W = N - o(N)$.

If the VERTEXCOVER instance is a NO instance, a vertex cover of $\bar{G}$ has size at least $(1 + \gamma)\bar{k}$. Consider an arbitrary seed set S' with $|S'| = k = \bar{m} + \bar{k}$. We can assume that S' does not contain any of $w_{u,v}^1, \dots, w_{u,v}^W$ in each gadget $G_{u,v}$, as seeding $e_{u,v}^u$ is strictly better than seeding any of $w_{u,v}^1, \dots, w_{u,v}^{\frac{W}{2}}$ and seeding $e_{u,v}^v$ is strictly better than seeding any of $w_{u,v}^{\frac{W}{2}+1}, \dots, w_{u,v}^W$ (if $e_{u,v}^u$ or $e_{u,v}^v$ is also seeded, we can remove the seeds from $w_{u,v}^1, \dots, w_{u,v}^W$ without changing the number of infected vertices). Moreover, we can assume that S' does not contain *both* $e_{u,v}^u$ and $e_{u,v}^v$ in each gadget $G_{u,v}$, as in this case either switching $e_{u,v}^u$ to u or switching $e_{u,v}^v$ to v will cause one more infection and thus strictly better: for example, $e_{u,v}^u$ will still be infected by u and $e_{u,v}^v$ if we seed u instead of $e_{u,v}^u$ (if either u or v is already seeded, we can just remove $e_{u,v}^u$ or $e_{u,v}^v$ from S' without changing the number of infected vertices).

We say that $G_{u,v}$ is *active* if we have either $u, e_{u,v}^u \in S'$ or $v, e_{u,v}^v \in S'$. From previous analyses, an active gadget contains at least $3 + W$ infected vertices. A gadget can fail to be active in the following two different ways:

Failure I none of $e_{u,v}^u, e_{u,v}^v$ is chosen in the seed set, and

Failure II none of u, v is chosen in the seed set.

Keeping our justified assumptions $\{e_{u,v}^u, e_{u,v}^v\} \not\subseteq S'$ and $w_{u,v}^1, \dots, w_{u,v}^W \notin S'$, by analyzing the cascade, it is straightforward to see that the expected number of infected vertices in $G_{u,v}$ is at most $\frac{3}{4}W + O(1)$ if one of the above two failures occur, and the number of infected vertices is 0 if both failures occur. To calculate an upper-bound of $\sigma_G(S')$, we only need to count the number of failures in all the gadgets, where the occurrences of both failures in a single gadget are counted as two failures. Since the occurrence of a single failure causes a loss of $\frac{1}{4}W + O(1)$ infected vertices and the occurrences of both failures in a gadget imply a loss of $W + O(1)$ infected vertices which is even more than $2 \times (\frac{1}{4}W + O(1))$, this way of counting only underestimates the number of uninfected vertices

Next, we aim to show that the total number of failures is at least $\gamma\bar{k}$. Let n_1, n_2 be the number of Failure I and II respectively. The number of seeds in the “inner parts” of all the $\bar{m}$ gadgets, $\bigcup_{u,v: (\bar{u}, \bar{v}) \in \bar{E}} \{e_{u,v}^u, e_{u,v}^v\}$, is $\bar{m} - n_1$, since exactly a seed is needed for one of $e_{u,v}^u, e_{u,v}^v$ in each $G_{u,v}$ to avoid a Failure I. As a result, the number of seeds allocated for those vertices representing vertices of $\bar{G}$, $\{v : \bar{v} \in \bar{V}\}$, is $k - (\bar{m} - n_1) = \bar{k} + n_1$. This symbolizes picking a set of $\bar{k} + n_1$ vertices in the VERTEXCOVER instance. By Theorem A.2, we need $\max\{\gamma\bar{k} - n_1, 0\}$ additional vertices to form a vertex cover. This implies that there are at least $\max\{\gamma\bar{k} - n_1, 0\}$ uncovered edges for the picked set of $\bar{k} + n_1$ vertices. Thus, $n_2 \geq \max\{\gamma\bar{k} - n_1, 0\}$. Therefore, the total number of failures is lower-bounded by

$$n_1 + n_2 \geq n_1 + \max\{\gamma\bar{k} - n_1, 0\} \geq \gamma\bar{k},$$

implying

$$\sigma_G(S') \leq N - (n_1 + n_2) \left(\frac{1}{4}W + O(1) \right) \leq \left(m - \frac{\gamma\bar{k}}{4} \right) W + o(W) \leq \left(1 - \frac{\gamma}{4d} \right) mW + o(W),$$

where the last inequality is due to our assumption $\bar{k} \geq \bar{m}/d$. We conclude the theorem by noticing $N = mW + o(W)$ and letting τ be a constant strictly less than $\frac{\gamma}{4d}$. $\square$

D REFINED UPPER BOUNDS FOR THE LINEAR THRESHOLD AND THE WEIGHTED INDEPENDENT CASCADE MODELS

The upper bounds in Theorem 3.1 and Theorem B.5 can be further refined, and the refined upper bounds will yield a local greedy heuristic with better performance than the heuristic based on upper bound in Theorem 3.1 and B.5.

THEOREM D.1. Given a uniform linear threshold model ULT_G with an undirected graph $G = (V, E)$ and a seed set $S \subseteq V$, we have

$$\sigma_G(S) \leq |S| + \sum_{v \in V \setminus S} \frac{\delta_v}{\deg(v)} (1 + \deg(v) - \delta_v),$$

where δ_v is the number of v 's neighbors in S .

THEOREM D.2. Given a weighted independent cascade model WIC_G with an undirected graph $G = (V, E)$ and a seed set $S \subseteq V$, we have

$$\sigma_G(S) \leq |S| + \sum_{v \in V \setminus S} \frac{\delta_v}{\deg(v)} (1 + \deg(v) - \delta_v),$$

where δ_v is the number of v 's neighbors in S .

Towards proving Theorem 3.1 and Theorem B.5, we have considered the lift of G with respect to S such that $\widehat{G}_S$ is a forest rooted at vertices in S , and we have shown that $\sigma_G(S) \leq \sigma_{\widehat{G}_S}(S)$ by coupling. We need a slightly more complicated way to “lift” G in order to show Theorem D.1, in which we make those vertices in ∂S to be the roots of the trees, where $\partial S = \{v \in V \setminus S \mid \exists s \in S : (s, v) \in E\}$ is the set of all non-seed vertices that are adjacent to a seed.

Similar to the definition of $\mathcal{P}_A$, let

$$\mathcal{P}_{\partial A} = \{P = ((v_1, v_2), (v_2, v_3), \dots, (v_{t-1}, v_t)) : v_1 \in \partial A; v_2, \dots, v_t \notin A; \forall i, j : v_i \neq v_j\}$$

be the set of all simple paths P that start from vertices in ∂A but never reach a vertex in A (we allow the paths passing a vertex in ∂A in the middle).

Definition D.3. Given an undirected graph $G = (V, E)$ and $A \subseteq V$, the *boundary lift* of G with respect to A , denoted by $\widehat{G}_A^b = (\widehat{V}, \widehat{E})$, is an undirected graph defined as follows.

- The vertex set is $\widehat{V} = A \cup \partial A \cup V_P$, where $V_P = \{v_P : P \in \mathcal{P}_{\partial A}\}$ is the set of vertices corresponding to the simple paths in $\mathcal{P}_{\partial A}$.
- For each $u \in A$ and $v \in \partial A$, include $(u, v) \in \widehat{E}$ if $(u, v) \in E$; for each $u \in \partial A$ and $v_P \in V_P$, include $(u, v_P) \in \widehat{E}$ if P is a path of length 1 that starts from u ; for each $v_{P_1}, v_{P_2} \in V_P$, include (v_{P_1}, v_{P_2}) if $|P_2| = |P_1| + 1$ and P_2, P_1 share the first $|P_1|$ common edges (or $|P_1| = |P_2| + 1$ and P_1, P_2 share the first $|P_2|$ common edges, since $\widehat{G}_A$ is undirected); do not include any edge between any two vertices in A ; do not include any edge between any two vertices in ∂A .

It is easy to see that the same vertex set $A \cup \partial A$ is in both $\widehat{G}_A^b$ and G . Moreover, for each $v \in \partial A$, if removing those edges from v to vertices in A , the connected component that contains v is a tree rooted at v . Figure 5 shows an example of G , $\widehat{G}_A$ and $\widehat{G}_A^b$.

Corresponding, the definitions for $\Psi(\cdot)$ and $\Phi(\cdot)$ are changed accordingly as follows.

$$\begin{aligned} \Psi(e) &= \begin{cases} \emptyset & \text{if } e = (u, v) \text{ for } u, v \in A \\ \{(u, v)\} & \text{if } e = (u, v) \text{ for } u \in A, v \in \partial A \\ \{(v_{P_1}, v_{P_2}) \mid P_2 = (P_1, e)\} & \text{Otherwise.} \end{cases} \\ \Phi(v) &= \begin{cases} \{v\} & \text{if } v \in A \cup \partial A \\ \{v_P \mid P \text{ ends at } v\} & \text{Otherwise.} \end{cases} \end{aligned}$$

It is straightforward to check that all the properties of the two functions in Sect. 4 continue to hold here.

Fig. 5. The lift and the boundary lift of G

PROOF OF THEOREM D.1. Following the same coupling arguments in the proof of Lemma 4.3, we can show that

$$\sigma_G(S) \leq \sigma_{\widehat{G}_S}(S).$$

For each $v \in \partial S$, we denote its neighbors as $s_1, \dots, s_{\delta_v}, u_1, \dots, u_{\deg(v)-\delta_v}$, where $s_1, \dots, s_{\delta_v}$ are v 's neighbors in S and $u_1, \dots, u_{\deg(v)-\delta_v}$ are v 's remaining neighbors. As we have mentioned, if we remove the δ_v edges $(s_1, v), \dots, (s_{\delta_v}, v)$, v is the root of a tree, with $u_1, \dots, u_{\deg(v)-\delta_v}$ being all the children of v . Since this tree contains no seed, we can apply the contraction argument in the proof of Lemma 4.2 to iteratively contract the leaves of this tree, until at the stage where the tree only contains v and $u_1, \dots, u_{\deg(v)-\delta_v}$. We do this contraction for each $v \in \partial S$. It is then simple to compute $\sigma_{\widehat{G}_S}(S)$: each $v \in \partial S$ is infected with probability $\frac{\delta_v}{\deg(v)}$, and it will further infect all its $\deg(v) - \delta_v$ children if infected. Therefore,

$$\sigma_{\widehat{G}_S}(S) = |S| + \sum_{v \in \partial S} \frac{\delta_v}{\deg(v)} (1 + \deg(v) - \delta_v).$$

The theorem follows from $\sigma_G(S) \leq \sigma_{\widehat{G}_S}(S)$. $\square$

PROOF OF THEOREM D.2. Following the same coupling arguments in the proof of Lemma B.4, we can show that

$$\sigma_G(S) \leq \sigma_{\widehat{G}_S^b}(S).$$

For each $v \in \partial S$, we denote its neighbors as $s_1, \dots, s_{\delta_v}, u_1, \dots, u_{\deg(v)-\delta_v}$, where $s_1, \dots, s_{\delta_v}$ are v 's neighbors in S and $u_1, \dots, u_{\deg(v)-\delta_v}$ are v 's remaining neighbors. By the same arguments in the previous proof, we can iteratively contract the leaves of this tree, until at the stage where the tree only contains v and $u_1, \dots, u_{\deg(v)-\delta_v}$. We do this contraction for each $v \in \partial S$. It is then simple to compute $\sigma_{\widehat{G}_S^b}(S)$: each $v \in \partial S$ is infected with probability $1 - \left(1 - \frac{1}{\deg(v)}\right)^{\delta_v} \leq \frac{\delta_v}{\deg(v)}$, and it will further infect all its $\deg(v) - \delta_v$ children if infected. Therefore,

$$\sigma_{\widehat{G}_S^b}(S) = |S| + \sum_{v \in \partial S} \frac{\delta_v}{\deg(v)} (1 + \deg(v) - \delta_v).$$

The theorem follows from $\sigma_G(S) \leq \sigma_{\widehat{G}_S^b}(S)$. $\square$

E EXPERIMENT RESULTS FOR INDEPENDENT CASCADE MODEL

E.1 The heuristic degreeDiscountIC

As we mentioned in Sect. 5.2, Chen et al. [6] proposed the heuristic degreeDiscountIC which is specifically designed for the uniform independent cascade model. The ideas and motivations are as follows. In degreeDiscount, the score of a candidate seed is computed by its degree minus the number of its neighbors that are already selected as seeds in the previous iterations. In other words, each edge connecting the candidate seed to the existing seed is “discounted” by 1. Intuitively, these edges, connecting to the existing seeds, play no role in further infections, so they should be discounted from the degree of the candidate seed. However, when considering the uniform independent cascade model especially with small parameter p , discounting by 1 is too much and thus inaccurate. The heuristic degreeDiscountIC shown in Algorithm 2 fixes this inaccuracy by using a more delicate estimation of this discount. Readers interested in more details of this heuristic can refer to [6].

Algorithm 2: degreeDiscountIC

```

input : An undirected graph  $G = (V, E)$ , parameter  $p \in [0, 1]$ , and an integer  $k$ 
initialize  $S = \emptyset$ ;
for each vertex  $v$  do
    compute its degree  $d_v$ ;
     $dd_v \leftarrow d_v$ ;
    initialize  $t_v \leftarrow 0$ ;
for  $i = 1$  to  $k$  do
    select  $u = \arg \max_v \{dd_v \mid v \in V \setminus S\}$ ;
     $S \leftarrow S \cup \{u\}$ ;
    for each neighbor  $v$  of  $u$  and  $v \in V \setminus S$  do
         $t_v \leftarrow t_v + 1$ ;
         $dd_v \leftarrow d_v - 2t_v - (d_v - t_v)t_vp$ ;
return  $S$ 

```

E.2 Results for the Independent Cascade Model

Weighted independent cascade model. The results for the weighted independent cascade model are shown in Fig. 6. Comparing Fig. 3 and Fig. 6, we can see that the shapes of the curves are very similar in each of the nine figures, although the numbers of infected vertices in both models are different (look at the values on the y -axis). This again suggests that the dynamics in both the linear threshold model and the weighted independent cascade model are similar.

Uniform independent cascade model. The results for the uniform independent cascade model are shown in Fig. 7. The parameter p is set to 0.01. Since degreeDiscountIC is specially designed for the uniform independent cascade model, we will only compare our heuristic to degreeDiscountIC, not degreeDiscount. Our heuristic has similar performance with degreeDiscountIC in all the six datasets. It is slightly worse than degreeDiscountIC on CA-GrQc, Nethept, DBLP, it matches the performance of degreeDiscountIC on ego-facebook and CA-HepPh, and it performs slightly better on com-YouTube. On average, the performance of our heuristic is 86.26% of the greedy algorithm (with 1,000,000 reverse reachable sets).

Fig. 6. Comparing the local greedy heuristic with $f_{vb}^{0.9}$ to the greedy maximum coverage algorithm on 1,000,000 reverse reachable sets (labelled as “greedy (RR sets)”), degreeDiscount, degreeDiscountIC, and the algorithm that simply pick k vertices with highest degrees (labelled as “maxDegree”). The cascade model is the weighted independent cascade model.

Fig. 7. Comparing the local greedy heuristic with $f_{vb}^{0.9}$ to the greedy maximum coverage algorithm on 1,000,000 reverse reachable sets (labelled as “greedy (RR sets)”), degreeDiscountIC, and the algorithm that simply pick k vertices with highest degrees (labelled as “maxDegree”). The cascade model is the uniform independent cascade model with $p = 0.01$.