Influence Maximization in Undirected Networks

Sanjeev Khanna[*] Brendan Lucier[†]

Abstract

We consider the problem of finding a set of k vertices of maximal total influence in a given *undirected* network, under the independent cascade (IC) model of influence spread. It is known that influence is submodular in the IC model, and hence a greedy algorithm achieves a $(1 - 1/e)$ approximation to this problem; moreover, it is known to be NP-hard to achieve a better approximation factor in directed networks.

We show that for undirected networks, this approximation barrier can be overcome: the greedy algorithm obtains an $(1 - 1/e + c)$ approximation to the set of optimal influence, for some constant $c > 0$. Our proof proceeds via probabilistic analysis of bond percolation in arbitrary finite networks. We also show that the influence maximization problems remains APX-hard in undirected networks.

[*] Dept. of Computer & Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA. sanjeev@cis.upenn.edu

[†] Microsoft Research New England, One Memorial Drive Cambridge, MA 02142. brlucier@microsoft.com

1 Introduction

Network diffusion models a scenario in which local interaction along edges in a graph can generate global cascades in network state. Such diffusion processes have attracted a significant amount of recent attention, having been studied in the context of network reliability [5, 3], bond percolation in statistical physics [14, 17], the spread of disease [6, 19], and diffusion of social influence [12, 13, 18].

We focus on the following standard diffusion process. We are given a (possibly directed) graph G with edge weights $p_e \in (0, 1]$. An unweighted graph H is then constructed at random as follows: independently for each edge e in G, we add e to graph H with probability p_e. We can therefore think of G as specifying a distribution over graphs, and think of H as a realization of a graph from this distribution. It is then assumed that diffusion (i.e., of disease, influence, spin state, etc.) spreads along the edges realized in H; that is, an infected vertex v will ultimately infect each node u reachable from v in H. This is known in the social influence literature as the *independent cascades* model of diffusion [12]. Despite being simple to describe, this random graph process has proven difficult to analyze in arbitrary networks. For example, it includes the Erdös-Renyi random graph model as a special case, and much of the prior research from percolation theory focuses on particular graph classes such as lattice networks [14, 17].

One striking example of an algorithmic problem that admits analysis for this diffusion model is the influence maximization problem. This problem is primarily motivated by applications to viral marketing: the goal is to select individuals in a network to target with a marketing intervention (e.g., free product samples) in order to maximize a subsequent cascade of product adoption. Formally, the algorithmic problem is as follows. For a set S of vertices of G, write $\mathrm{val}(S)$ for the expected number of nodes reachable from vertices in S in the realized graph H. In the context of influence spread, we can think of $\mathrm{val}(S)$ as a measure of the influence of set S. Given network G, the *influence maximization problem* is to find a set S of size k such that $\mathrm{val}(S)$ is maximized.

Kempe et al. [12] first formulated this problem, and noted that the function $\mathrm{val}(S)$ is monotone and submodular. The influence maximization problem is therefore an instance of monotone submodular function maximization subject to a cardinality constraint, and hence the greedy algorithm (which repeatedly selects the node that maximizes the marginal contribution to $\mathrm{val}(\cdot)$) obtains a $(1 - 1/e)$ approximation in polynomial time. Moreover, this approximation factor is the best possible: it is NP-hard to achieve an approximation $(1 - 1/e + \epsilon)$ for any $\epsilon > 0$ [13].

The above results apply to general *directed* networks. However, many network diffusion processes are modeled on undirected networks. For instance, one might estimate the probability of one individual influencing another as a function of the amount of interaction or contact between them, which is symmetric. We ask: how well can the influence maximization problem be approximated *in undirected networks*? Note that the directed-network analysis of the greedy algorithm implies that one can achieve at least a $(1 - 1/e)$-approximation, so the relevant question is whether the lower bound of $(1 - 1/e)$ for submodular function maximization applies to influence maximization in undirected networks. We show that the answer is *no*: there exists a constant $c > 0$ such that it is possible to approximate the maximum influence to within a factor of $(1 - 1/e + c)$ in polynomial time. Moreover, this approximation factor is achieved by the standard greedy algorithm.

Our Results and Techniques. Our main result is that the greedy algorithm for submodular function maximization achieves an approximation factor of $(1-1/e+c)$ for the influence maximization problem in undirected networks, for some fixed $c > 0$. This result applies even to a version of the influence maximization problem that is generalized in two ways: first, each node v has a non-negative weight $w(v)$ and the goal is to maximize the total weight of nodes influenced; second, there is a specified set U of permitted vertices and the selected set S must be contained in U.

We do not make a serious attempt to optimize the constant c; we expect that our analysis can

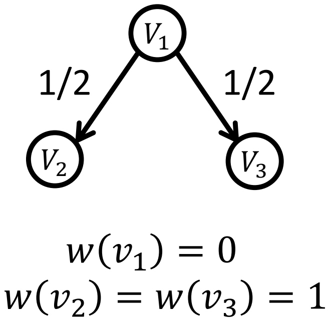


Figure 1: A tight example for the performance of greedy in a directed network, for $k = 2$. The optimal solution is $\{v_2, v_3\}$ for a total weighted influence of 2. The greedy algorithm could select node v_1 first, then v_2, for a total weighted influence of $\frac{3}{2}$.

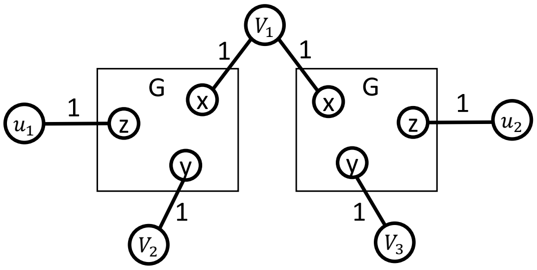
be tightened to achieve a significantly improved constant, but we leave this for future work. As a complement to our main result, we also show that the influence maximization problems remains APX-hard in undirected networks. In what follows, we give an overview of the approach used in establishing our main result.

Before describing our techniques in detail, let us first consider a few relevant examples. To illustrate the crucial difference between directed and undirected networks, consider the case $k = 2$. In Figure 1 we describe a directed network for which the approximation factor of the greedy algorithm is 3/4, which is tight for $k = 2$. The example is a tree with edges directed away from the root, where the root v_1 has two children v_2 and v_3; the weight of v_1 is 0, and the weight of each child is 1. Each edge has weight 1/2. In this example, the optimal solution is $\{v_2, v_3\}$, with $\text{val}(\{v_2, v_3\}) = 2$. The greedy algorithm might[1] first select node v_1, then either v_2 or v_3, for a total value of $\frac{3}{2}$. The crucial feature driving the gap in this example is that the nodes in the optimal solution influence disjoint sets of nodes, but there is a node v_1 whose influence has a high amount of intersection with that of v_2 and of v_3 (over randomness in the influence process).

Consider what occurs in this example if edges are undirected. The optimal solution is still $\{v_2, v_3\}$, and $\text{val}(\{v_2, v_3\}) = 2$ as before. However, we now have $\text{val}(\{v_2\}) = \text{val}(\{v_3\}) = \frac{5}{4}$ (due to the path through v that connects v_2 and v_3) whereas $\text{val}(\{v_1\}) = 1$. The greedy algorithm therefore finds the optimal solution in the undirected version of the example. Intuitively, this difference is driven by the fact that, in undirected networks, the presence of a node whose range of influence overlaps that of multiple elements of the optimal solution, there must be overlapping influence among the optimal elements themselves. The key to our main result is showing that this intuition applies more generally in undirected networks.

Establishing the general result requires that we overcome subtle difficulties in the analysis of percolation on networks. For instance, our analysis requires that we establish the following fact (which we refer to as the XYZ lemma, and which appears below as Lemma 6): for any three nodes x, y, and z in G, if all three nodes are connected (i.e., are in the same component) with some probability p, then at least one pair of these nodes must be connected with probability significantly larger than p. To build some intuition for why this is necessary for our result, suppose there exists a graph G and vertices x, y, z that are all connected with probability ϵ, but the probability that any two are connected without the other is significantly less than ϵ. Using such a graph G as a building block, we can construct a counter-example to our main result. See Figure 2, where we show

[1] Note that one could perturb the vertex weights so that this is the unique outcome of the greedy algorithm.



$w(v_1) = 0,\ w(v_2) = w(v_3) = 1,\ w(u_1) = w(u_2) = {}^1/_\epsilon$

Figure 2: An example illustrating that the XYZ lemma is necessary to show that the performance of the greedy algorithm improves on undirected networks. The set of valid seed nodes is $U = \{v_1, v_2, v_3\}$, and $k = 2$. The supposed graph G is such that x,y,z are all connected with probability $\epsilon > 0$, but the probability that exactly two of these nodes are connected is negligible. The optimal solution is $\{v_2, v_3\}$ for a total weighted influence of $4 + o(\epsilon)$. The greedy algorithm selects v_1 first, then v_2, for a total weighted influence of $3 + O(\epsilon)$.

how to construct (using such a G) an undirected network for which the approximation factor of the greedy algorithm is arbitrarily close to 3/4, when $k = 2$. This example extends to larger k as well, resulting in an approximation factor arbitrarily close to $1 - 1/e$. Our main result therefore requires, in particular, that we prove the XYZ lemma (in addition to other technical facts about percolation on undirected networks).

We now give a more detailed overview of our approach. Our analysis of the greedy algorithm proceeds in three steps. First, we establish that the worst-case instances of the influence maximization problem must be of a particular form. Specifically, the optimal solution, $\text{OPT} = \{o_1, \ldots, o_k\}$, should be such that the influence function $\text{val}(\cdot)$ is nearly linear on OPT, and nearly equal for each singleton in OPT. We show that if this were not the case, then we can immediately establish that the greedy algorithm will achieve an approximation factor significantly better than $(1 - 1/e)$. This part of the analysis does not make use of undirectedness.

In the second step, we consider the implications of the linearity of $\text{val}(\cdot)$ on OPT when the network is undirected. In particular, it must be that any given node in OPT is very rarely in the same component as another node from OPT, over realizations of the network. Indeed, if this occurred often for some $o_i \in \text{OPT}$, then $\text{val}(o_i)$ would be significantly smaller than $\text{val}(\text{OPT}) - \text{val}(\text{OPT} \backslash \{o_i\})$, the marginal value of o_i given the other elements of OPT, violating approximate linearity.

As it turns out, this non-connectedness of OPT has strong implications for the relationship between OPT and the set of nodes selected by the greedy algorithm. Consider the first $k/4$ nodes selected by the greedy algorithm, say S. If the greedy algorithm is to have a low approximation factor relative to OPT, then it should be that $\text{val}(o_i)$ is significantly larger than $\text{val}(o_i \cup S) - \text{val}(S)$; otherwise, the greedy algorithm could select o_i as its next element and achieve a better-than-expected approximation factor. Intuitively, in order for $\text{val}(o_i \cup S) - \text{val}(S)$ to be small relative to $\text{val}(o_i)$, it should be that o_i is often in the same component as some node in S. But, since there are k elements of OPT and $k/4$ elements of S, we arrive at a contradiction: if each o_i is often in the same component as a node in S, then a pigeonhole argument implies that some nodes in OPT must often be in the same component as other nodes from OPT, which we know does not occur.

The above argument relies on the intuition that, in order for $\text{val}(o_i \cup S) - \text{val}(S)$ to be small (relative to $\text{val}(o_i)$), it must be that o_i is often in the same component as some node in S. Formalizing this intuition is the third step of our analysis, and the most technical. To see why this is not obvious, suppose that there is some node $s \in S$ such that o_i and s are very rarely in the same component but, when they are, that component is extremely large. Otherwise, when o_i and s are in different components, those components are very small. In this hypothetical situation, a large fraction of $\mathrm{val}(o_i)$ is due to events in which o_i is in the same component as s, and hence $\mathrm{val}(o_i \cup S) - \mathrm{val}(S)$ is small. We must therefore prove that this scenario cannot occur: if the component containing o_i is large conditional on it including s, then one of o_i or s must have a large expected component size unconditionally. Establishing this fact, which may be of independent interest, requires a technical probabilistic analysis of the random graph model. This analysis is captured in our XYZ Lemma (Lemma 6), which relates conditional and unconditional connection probabilities.

Related work Models of network influence have long been studied in the sociology and marketing literature [11, 21, 8]. The problem of finding the most influential set of nodes in a network was originally posed by Domingos and Richardson [7, 20]. A formal development of the influence maximization problem and the independent cascades model, along with a greedy algorithm based upon submodular maximization, was given by Kempe et al. [12]. The lower bound of $(1 - 1/e)$ on approximability of this problem was subsequently established in Kempe at al. [13]. Many subsequent works have studied the nature of diffusion in online social networks, using empirical data to estimate influence probabilities and infer network topology; see [16, 9, 15].

It is known that many alternative formulations of the influence maximization problem are computationally difficult. In particular, the problem of determining influence spread given a particular seed set, in the IC model, is #P-hard [4].

Various other models of influence spread have been proposed and analyzed in the literature, with much of the prior work focusing on models that admit submodular influence functions [18, 13]. Such models have also been extended to include interations between multiple diffusive processes [10, 2]. We focus on the IC model, and leave open the question of whether these alternative models also admit improved approximation factors in undirected networks.

2 Preliminaries

The input to the undirected influence maximization problem is a five-tuple $\langle G(V, E), U, p, w, k\rangle$ where $G(V, E)$ is an undirected graph, $U \subseteq V$ is the set of allowed seed vertices, $p : E \to [0, 1]$ is a probability function on the edges, w is a non-negative integer weight function on the vertices, and k is a positive integer. A problem instance defines an influence function, $\mathrm{val}\colon 2^V \to \mathbb{R}_+$, in the following manner. First, we define a distribution over unweighted, undirected graphs $H(V', E')$ as follows: $V' = V$, and for each $e \in E$ independently, we add e to E' with probability p_e. Then, for any $S \subseteq V$, we define $\mathrm{val}(S)$ to be the expectation, over realizations of graph H, of the total weight of all vertices that lie in the same component as a node in S.

The goal of the *influence maximization problem* is to choose a set of seed vertices $S \subseteq U$, with $|S| = k$, such that $\mathrm{val}(S)$ is maximized. We will tend to write $\mathrm{val}(S \mid T)$ to mean $\mathrm{val}(S \cup T) - \mathrm{val}(T)$, the marginal value of S given T. The *greedy algorithm* for the influence maximization problem proceeds by repeatedly adding to its solution set S the vertex $v \in U$ that maximizes $\mathrm{val}(v \mid S)$, until k nodes have been chosen. It is known that val is a monotone submodular function [12], and hence the greedy algorithm obtains a $(1 - 1/e)$-approximation to the influence maximization problem.

Given an instance of the undirected influence maximization problem, we will write GRD for the solution returned by the greedy algorithm, and $\mathrm{OPT} = \{O_1, \dots, O_k\}$ for the optimal solution. Also, given sets of vertices S and T, we will tend to write $S \to T$ for the event that some vertex in S lies in the same component as a vertex in T, over realizations of the random graph process.

In the remainder of the paper, we will assume without loss of generality that the given instance is unweighted, that is, each vertex in $G(V, E)$ has unit weight. A weighted instance can be transformed in polynomial-time into an unweighted instance such that he reduction preserves the approximation factor as follows. Suppose we are given a weighted instance $G(V,E)$ with a weight function w on the vertices. Let $\Gamma = \Theta(\text{val}(\text{OPT}))$ be some estimate of the value of OPT that is accurate to within a constant factor (such an estimate can be computed by simply running the greedy strategy and invoking the standard analysis of the greedy algorithm). As a first step in our transformation, we create a weighted graph $G'(V,E)$ such that each vertex v in G' has a positive integer weight $w'(v)$ that is in the range 1 throughout $\Theta(n^2)$. We do so by defining the weight function $w'(v) = \lfloor w(v)/(\Gamma/n^2)\rfloor + 1$. Clearly, $w'(v) \geq 1$ for all each vertex v in G. Furthermore, since no vertex in G can have weight greater than $\Theta(\Gamma)$, the upper bound of $\Theta(n^2)$ on the weight of any vertex in G' follows. It is easy to verify that for any set S of seed vertices, we have

$$(\Gamma/n^2)\text{val}_{G'}(S) - n(\Gamma/n^2) \leq \text{val}_G(S) \leq (\Gamma/n^2)\text{val}_{G'}(S).$$

Thus any α-approximate solution in G' can be mapped to an $\alpha(1-o(1))$-approximate solution in G. As a next and the final step in the transformation, we convert $G'(V,E)$ into an unweighted graph $G''(V'',E'')$ such that for any set S of seed vertices, we have $\text{val}_{G'}(S) = \text{val}_{G''}(S)$, thus completing the proof. This is done by simply attaching $(w'(v)-1)$ auxiliary vertices to each vertex v in G', such that these auxiliary vertices are connected by edges with probability 1. Hence whenever vertex v is reached in G'', a total weight of $w'(v)$ is collected – same as in the graph G'. The set of allowed seed vertices is kept unchanged through both transformations above. It is easy to verify that this transformation can be done in polynomial-time since all vertex weights are bounded by $\Theta(n^2)$.

3 Main Result: Approximation Factor of the Greedy Algorithm

In this section we prove our main result, which is a bound on the approximation factor of the greedy algorithm for the undirected influence maximization problem.

Theorem 1. *There exists a constant $c > 0$ such that, for any instance $\langle G(V,E), U, p, w, k\rangle$ of the undirected influence maximization problem,* $\text{val}(\text{GRD}) > (1 - \frac{1}{e} + c)\text{val}(\text{OPT})$.

Our proof of Theorem 1 proceeds as follows. In Section 3.1 we show that Theorem 1 is true whenever OPT is not of a particular "balanced" form (Lemma 1 and Lemma 2). Then, in Section 3.2, we consider the state of the greedy algorithm after having selected $k/4$ nodes; we show that either the greedy algorithm can select its next vertex to have a very high marginal value, proving Theorem 1 (Lemma 4), or else the balanced form of OPT leads to a contradiction. This final contradiction requires that we establish a number of technical properties of the undirected random graph process, which are proved in Sections 4, 5, and B.

3.1 Reduction to Balanced Optimal Instances

For a subset X of vertices, we define the *normalized influence of* X, denoted by $\rho(X)$, to be the ratio

$$\frac{\text{val}(X)/|X|}{\text{val}(\text{OPT})/k}.$$

Given an $\epsilon > 0$, we say that a set X of vertices is *ϵ-uniform* if for each vertex $x \in X$, $(1-\epsilon) \leq \rho(x|X \setminus \{x\}) \leq (1+\epsilon)$, and that X is *ϵ-independent* if for each vertex $x \in X$, $\Pr[x \to X \setminus \{x\}] \leq \epsilon$. We refer to a set X as *ϵ-balanced* if it is both ϵ-uniform and ϵ-independent. Roughly speaking, being ϵ-balanced means that $\text{val}(\cdot)$ is approximately linear and uniform on X. Our goal in this section is to show that for any $\epsilon > 0$, either the performance of the greedy algorithm is at least $(1-1/e+f(\epsilon))$ for some positive function $f(\epsilon)$, or the optimal solution is ϵ-balanced for some $\epsilon > 0$. In the subsequent sections, we will show our main result, namely, the greedy achieves a strictly better than $(1-1/e)$ performance ratio whenever OPT is ϵ-balanced for some small enough $\epsilon > 0$.

We start with the following simple lemma: whenever there exists a set X of vertices whose normalized influence is strictly greater than 1 and $\text{val}(X) = \Omega(\text{val}(\text{OPT}))$, then the greedy algorithm beats the $1 - 1/e$ performance ratio. The proof proceeds by modifying the standard analysis of the greedy algorithm to first compare its performance against $\text{val}(X)$ for some number of iterations, then against $\text{val}(\text{OPT})$ for the remaining iterations.

Lemma 1. *Suppose for some $\epsilon > 0$ and $\delta \in (0, 1)$ there exists a set $X \subseteq \text{OPT}$ with $|X| = \delta k$ such that $\rho(X) > (1 + \epsilon)$. Then*

$$\text{val}(\text{GRD}) > \left(1 - \frac{1}{e}\left(1 - \frac{\delta^2\epsilon^2}{4}\right)\right) \text{val}(\text{OPT}).$$

Proof. Let $\gamma \in (0, 1)$ be a constant to be determined later. We analyze the performance of the greedy algorithm by comparing the value of the first $\gamma\delta k$ sets relative to the sets in the collection X, and then the rest relative to the residual value of the optimal. The value of the first $\gamma\delta k$ sets chosen by the greedy algorithm, say a collection Y, is at least

$$\text{val}(Y) \geq \text{val}(X)\left(1 - \frac{1}{e^\gamma}\right) > \frac{(1+\epsilon)\cdot\delta\cdot\text{val}(\text{OPT})}{k}\left(1 - \frac{1}{e^\gamma}\right),$$

where the first inequality follows by applying the standard greedy analysis and measuring $\text{val}(Y)$ relative to $\text{val}(X)$, and the second inequality follows from our assumption about the set X.

The next $(1 - \gamma\delta)k$ sets chosen by the greedy algorithm, say a collection Z, gets an additional contribution of at least

$$\text{val}(Z) \geq (\text{val}(\text{OPT}) - \text{val}(Y))\left[1 - \frac{1}{e^{1-\gamma\delta}}\right].$$

Thus total value of the sets chosen by the greedy must be at least

$$\text{val}(Y) + \text{val}(Z) \geq \text{val}(\text{OPT}) - \frac{\text{val}(\text{OPT}) - \text{val}(Y)}{e^{1-\gamma\delta}}.$$

Hence the deficit of the greedy algorithm with respect to $\text{val}(\text{OPT})$ can be bounded by

$$\begin{aligned} \frac{\text{val}(\text{OPT}) - \text{val}(Y)}{e^{1-\gamma\delta}} &\leq \frac{\text{val}(\text{OPT}) - (1+\epsilon)\delta\text{val}(\text{OPT})(1 - \frac{1}{e^\gamma})}{e^{1-\gamma\delta}} \\ &\leq \frac{\text{val}(\text{OPT})}{e}\left[e^{\gamma\delta}(1 - (1+\epsilon)\delta) + e^{-\gamma(1-\delta)}(1+\epsilon)\delta\right] \end{aligned}$$

Let $\Gamma = \left[e^{\gamma\delta}(1 - (1+\epsilon)\delta) + e^{-\gamma(1-\delta)}(1+\epsilon)\delta\right]$. To complete the proof of the lemma, it suffices to show that by choosing $\gamma = (\delta\epsilon)/2$, we can bound Γ by $\left(1 - \frac{\delta^2\epsilon^2}{4}\right)$, that is, $\Gamma < 1$ for ϵ, δ are both positive. Using the fact that $e^x \leq 1 + x + x^2$ for $|x| < 1$, we have

$$\begin{aligned} \Gamma &\leq (1 + \gamma\delta + \gamma^2\delta^2)(1+\epsilon)\delta(1 - (1+\epsilon)\delta) + (1 - \gamma + \gamma\delta + \gamma^2(1-\delta)^2) \\ &\leq 1 + \gamma\delta - \gamma(1+\epsilon)\delta + \gamma^2(\delta^2 - \delta^3(1+\epsilon) + (1-\delta)^2) \\ &\leq 1 - \gamma\delta\epsilon + \gamma^2(\delta^2 + (1-\delta)^2) \\ &\leq 1 - \gamma\delta\epsilon + \gamma^2 \leq 1 - \frac{\delta^2\epsilon^2}{4}, \end{aligned}$$

where the last but one inequality follows from the fact that $(\delta^2 + (1-\delta)^2) \leq 1$ for $\delta \in (0, 1)$, and the last equality follows by choosing $\gamma = (\delta\epsilon)/2$. The assertion of the lemma thus follows. □

Our next goal is to show that given any optimal solution OPT, either it contains a set X of vertices with $\rho(X) > 1$ and $\mathrm{val}(X) = \Omega(\mathrm{val}(\mathrm{OPT}))$ (and thus Lemma 1 ensures that greedy performs strictly better than $1 - 1/e$) or OPT is essentially ϵ-uniform. The proof proceeds by showing that if OPT contains many vertices whose value is much larger or smaller than $\mathrm{val}(\mathrm{OPT})/k$, then this implies the existence of a set X satisfying the conditions of Lemma 1. One subtlety is that it is not actually enough for many vertices to have value larger than $\mathrm{val}(\mathrm{OPT})/k$; what we require is that there are many such large-value vertices even if we only consider the marginal contributions given some small, fixed subset of OPT. This is what motivates the focus on marginal values given H in the statement of Lemma 2.

Lemma 2. *For any $\epsilon > 0$, either* OPT *contains a set X of vertices whose normalized influence is strictly greater than* 1 *and* $\mathrm{val}(X) = \Omega(\mathrm{val}(\mathrm{OPT}))$, *or* OPT *can be partitioned into three sets L, M, and H such that (a)* $\mathrm{val}(H) \leq \epsilon^2 \cdot \mathrm{val}(\mathrm{OPT})$, *(b)* $|M| \geq (1-2\epsilon)k$, *and (c) each $O_i \in M$ satisfies*

$$\frac{(1-\epsilon)\mathrm{val}(\mathrm{OPT})}{k} \leq \mathrm{val}(O_i|O_{-i} \cup H) \leq \mathrm{val}(O_i|H) \leq \frac{(1+\epsilon)\mathrm{val}(\mathrm{OPT})}{k},$$

where O_{-i} denotes the set $M \setminus \{O_i\}$.

Proof. We will give an iterative procedure to construct the decomposition into sets L, M and H as above, and show that the procedure succeeds unless OPT contains a set X of vertices whose normalized influence is strictly greater than 1 and $\mathrm{val}(X) = \Omega(\mathrm{val}(\mathrm{OPT}))$.

Initialize Z to contain the vertices $\{O_1, ..., O_k\}$ in OPT, and initialize $L = \emptyset$. While there exists a vertex $O_i \in Z$ such that $\mathrm{val}(O_i|Z \setminus O_i) < \frac{(1-\epsilon)\mathrm{val}(\mathrm{OPT})}{k}$, do $L = L \cup \{O_i\}$ and $Z = Z \setminus \{O_i\}$. If upon termination, the set L contains more than ϵk vertices, then the set $X = \mathrm{OPT} \setminus L$ satisfies

$$\begin{aligned}
\mathrm{val}(X) &\geq \mathrm{val}(\mathrm{OPT}) - |L| \cdot \frac{(1-\epsilon)\mathrm{val}(\mathrm{OPT})}{k} \\
&\geq \epsilon\mathrm{val}(\mathrm{OPT}) + (k - |L|)\frac{(1-\epsilon)\mathrm{val}(\mathrm{OPT})}{k} \\
&\geq \epsilon\mathrm{val}(\mathrm{OPT}) + |X|\frac{(1-\epsilon)\mathrm{val}(\mathrm{OPT})}{k} \\
&\geq \epsilon\frac{|L| + |X|}{k}\mathrm{val}(\mathrm{OPT}) + |X|\frac{(1-\epsilon)\mathrm{val}(\mathrm{OPT})}{k} \\
&\geq \epsilon\frac{|L|}{k}\mathrm{val}(\mathrm{OPT}) + |X|\frac{\mathrm{val}(\mathrm{OPT})}{k} \\
&\geq \epsilon^2\mathrm{val}(\mathrm{OPT}) + |X|\frac{\mathrm{val}(\mathrm{OPT})}{k}.
\end{aligned}$$

This gives us the desired set X with normalized influence strictly greater than 1 and $\mathrm{val}(X) = \Omega(\mathrm{val}(\mathrm{OPT}))$. Assuming $|L| < \epsilon k$, we continue with the decomposition process on the remaining set Z to identify the sets M and H. Note that in this case, we know that $\mathrm{val}(L) \leq \epsilon \cdot \mathrm{val}(\mathrm{OPT})$. Let σ be an ordering of the vertices in Z created in the following manner. We choose $O_{\sigma(1)}$ to be a vertex $O \in Z$ that maximizes $\mathrm{val}(O)$. Then $O_{\sigma(2)}$ is chosen to be a vertex $O \in Z$ that maximizes $\mathrm{val}(O|O_{\sigma(1)})$. In general, we choose $O_{\sigma(2)}$ is chosen to be a vertex $O \in Z$ that maximizes $\mathrm{val}(O|O_{\sigma(1)}, \ldots, O_{\sigma(i-1)})$. Now consider the largest index j such that

$$\mathrm{val}(O_{\sigma(j)}|O_{\sigma(1)}, \ldots, O_{\sigma(j-1)}) > \frac{(1+\epsilon)\mathrm{val}(\mathrm{OPT})}{k}.$$

If $\mathrm{val}(O_{\sigma(1)} \cup O_{\sigma(2)} \cup \ldots \cup O_{\sigma(j)}) \geq \epsilon^2\mathrm{OPT}$, then $X = \{O_{\sigma(1)}, \ldots, O_{\sigma(j)}\}$ gives us a set with normalized influence strictly greater than 1 and $\mathrm{val}(X) = \Omega(\mathrm{val}(\mathrm{OPT}))$. Otherwise, we continue with our decomposition and define $H = \{O_{\sigma(1)}, \ldots, O_{\sigma(j)}\}$, and $M = Z \setminus H$. We now argue that for each $O_i \in \mathrm{OPT}'$, we have

$$\frac{(1-\epsilon)\mathrm{val}(\mathrm{OPT})}{k} \leq \mathrm{val}(O_i | O_{-i}, H) \leq \mathrm{val}(O_i | H) \leq \frac{(1+\epsilon)\mathrm{val}(\mathrm{OPT})}{k}.$$

The first inequality follows from the fact that $O_i \notin L$. The second inequality follows from submodularity of the influence function. The last inequality follows because $O_i \notin H$. This completes the proof of Lemma 2. □

If OPT contains a subset X with $\rho(X) > 1$ and $\mathrm{val}(X) = \Omega(\mathrm{val}(\mathrm{OPT}))$, then by Lemma 1 we are already done. So we assume from here onwards that OPT does not contain the desired set of normalized influence strictly greater than 1 and thus admits a decomposition into sets L, M, and H as outlined in Lemma 2. We next show that M contains a subset M' of size at least $|M| - \epsilon k$ such that M' is (5ϵ)-independent.

Lemma 3. *Let L, M, H constitute a decomposition of* OPT *satisfying the properties of Lemma 2. Then for any $\epsilon \in (0, 1/3)$, there exists a subset $M' \subseteq M$ of size at least $|M| - \epsilon k$ such that for each $O_i \in M'$, we have* $\Pr[O_i \to O_{-i}] \leq 5\epsilon$, *where we define the set O_{-i} to be $M' \setminus \{O_i\}$.*

Proof. Let $M_1 = \{O_i \in M \mid \Pr[O_i \to H] > 2\epsilon\}$. We will first show that $|M_1| \leq \epsilon k$. To see this, note that

$$\begin{aligned}
\mathrm{val}(H) &\geq \sum_{O_i \in M} \Pr[O_i \to H]\mathrm{val}(O_i \mid M \setminus \{O_i\}) \\
&\geq 2\epsilon \cdot \sum_{O_i \in M_1} \Pr[O_i \to H]\mathrm{val}(O_i \mid M \setminus \{O_i\}) \\
&\geq 2\epsilon \cdot \sum_{O_i \in M_1} \Pr[O_i \to H]\mathrm{val}(O_i \mid M \setminus \{O_i\}, H) \\
&\geq 2\epsilon |M_1| \frac{(1-\epsilon)\mathrm{val}(\mathrm{OPT})}{k} \\
&> \epsilon^2 \cdot \mathrm{val}(\mathrm{OPT})
\end{aligned}$$

for any $\epsilon \in (0, 1/3)$ whenever $|M_1| \geq \epsilon k$. But this contradicts our assumption that $\mathrm{val}(H) \leq \epsilon^2 \mathrm{val}(\mathrm{OPT})$.

Let $M' = M \setminus M_1$. We will now show that for each vertex $O_i \in M'$, we have $\Pr[O_i \to M' \setminus \{O_i\}]$ is at most 5ϵ. We argue this as follows:

$$\begin{aligned}
\Pr[O_i \to M' \setminus \{O_i\}] &= \Pr[O_i \to M' \setminus \{O_i\} \mid O_i \to H] \cdot \Pr[O_i \to H] \\
&\quad + \Pr[O_i \to M' \setminus \{O_i\} \mid O_i \not\to H] \cdot \Pr[O_i \not\to H] \\
&\leq \Pr[O_i \to H] + \Pr[O_i \to M' \setminus \{O_i\} \mid O_i \not\to H]
\end{aligned}$$

Thus it suffices to show that $\Pr[O_i \to M' \setminus \{O_i\} \mid O_i \not\to H] \leq 3\epsilon$ for all $O_i \in M'$. To see this, we first observe that for each $O_i \in M'$ (in fact the analysis below applies to each $O_i \in M$ and not just M'), we have

$$\sum_j \Pr[O_i \to j \wedge O_i \to O_{-i} \wedge O_i \not\to H] \leq \frac{2\epsilon \mathrm{val}(\mathrm{OPT})}{k},$$

since $\mathrm{val}(O_i | O_{-i}, H) \geq (1-\epsilon)\frac{\mathrm{val}(\mathrm{OPT})}{k}$ and $\mathrm{val}(O_i | H) \leq (1+\epsilon)\frac{\mathrm{val}(\mathrm{OPT})}{k}$. So, we have

$$\begin{aligned}\frac{2\epsilon\mathrm{val}(\mathrm{OPT})}{k} &\geq \sum_j \Pr[O_i \to j \wedge O_i \to O_{-i} \wedge O_i \not\to H] \\ &= \sum_j \Pr[O_i \to O_{-i} \mid O_i \to j \wedge O_i \not\to H] \cdot \Pr[O_i \to j \wedge O_i \not\to H] \\ &\geq \sum_j \Pr[O_i \to O_{-i} \mid O_i \not\to H] \cdot \Pr[O_i \to j \wedge O_i \not\to H] \\ &\geq \Pr[O_i \to O_{-i} \mid O_i \not\to H] \cdot \mathrm{val}(O_i \mid H) \\ &\geq \Pr[O_i \to O_{-i} \mid O_i \not\to H] \cdot (1-\epsilon)\frac{\mathrm{val}(\mathrm{OPT})}{k}\end{aligned}$$

Hence it follows that

$$\Pr[O_i \to O_{-i} \mid O_i \not\to H] \leq \frac{2\epsilon}{1-\epsilon} \leq 3\epsilon,$$

for any $\epsilon \in (0, 1/3)$, concluding the proof of the lemma.

□

3.2 Proving Theorem 1 for Balanced Optimal Instances

Let L, M, M', H be a decomposition of OPT satisfying the properties of Lemma 2 and Lemma 3. Let $S = \{g_1, g_2, ..., g_{k/4}\}$ be the first $k/4$ nodes selected by the greedy algorithm.

The strategy of the proof is as follows. We first show that if the greedy algorithm does not achieve an approximation much better than $1 - 1/e$, then the marginal influence of each $O_i \in M'$, given S, must not be too large (Lemma 4). On the other hand, since elements of M' have low probability of being in the same connected component, we must conclude that many elements of M' have low probability of being in the same connected component as S (Lemma 5). These two facts are seemingly contradictory, since the marginal influence of O_i given S is low when, roughly speaking, the probability that O_i and S are in the same component is large. To formalize this intuition, we must establish bounds on the correlation between component sizes and connectivity events; this bound is captured in the XYZ Lemma (Lemma 6). Finally, to apply the XYZ Lemma to sets M' and S, we will show that it suffices to consider the part of each set's influence that is "well-behaved" in a certain sense (Lemma 7). We begin by establishing that if $\mathrm{val}(O_i \mid S)$ is too large for any $O_i \in M'$, then the greedy algorithm attains an approximation factor better than $1 - 1/e$.

Lemma 4. *Suppose there exists an $O_i \in M'$ such that $v(O_i \mid S) \geq \frac{4}{5} \cdot \frac{\mathrm{val}(\mathrm{OPT})}{k}$. Then $\mathrm{val}(GRD) > (1 - \frac{1}{e} + c)\mathrm{val}(OPT)$ for a fixed constant c.*

Proof. Suppose $v(O_i \mid S) \geq \frac{4}{5} \cdot \frac{\mathrm{val}(\mathrm{OPT})}{k}$. Then O_i is a candidate for $g_{k/4+1}$ (the greedy element to pick after S), and in particular all previous greedy elements from S achieved at least this much marginal value. But since $\frac{4}{5} > (1-1/k)^{k/4}$ for all $k \geq 4$, we conclude that $\mathrm{val}(S) > (1-(1-1/k)^{k/4} + c')\mathrm{val}(\mathrm{OPT})$ for some fixed constant c'. We then have $\mathrm{val}(\mathrm{GRD}) > (1 - \frac{1}{e} + c'(1-1/k)^{3k/4})\mathrm{val}(\mathrm{OPT})$ as required, where the value of c in the lemma statement is $c' \cdot e^{-3/4}$. □

We next establish that, since $\Pr[O_i \to O_{-i}] \leq 5\epsilon$ for each $O_i \in M'$, it must be that many $O_i \in M'$ have low probability of sharing a component with a node in S. We prove Lemma 5 in Section 4.

Lemma 5. *There exists $M'' \subseteq M'$ with $|M''| \geq k/3$ such that $\Pr[O_i \to S] < 14\sqrt{\epsilon}$ for all $O_i \in M''$.*

Our goal is to show that Lemma 4 and Lemma 5 together imply a contradiction. The following lemma, which we call the XYZ Lemma, bounds the extent of correlation between component sizes in bond percolation, and is crucial to our result. Its proof appears in Section 5.

Lemma 6 (XYZ Lemma)**.** *In any undirected graph G with a probability function on the edges, For any 3 vertices x, y, and z, we have* $\Pr[y \to z] \geq \frac{1}{4} \cdot \Pr[x \to z \mid x \to y] \cdot \Pr[x \to y \mid x \to z]$.

Finally, we argue that a large fraction of the total influence of M'' is captured by events in which M'' influences a node j in the following well-behaved way. First, the probability that M'' influences j is approxiately the sum of the probabilities that each $O_i \in M''$ influences j. Second, the probability that both M'' and S influence j is not too much smaller than the probability that M'' influences j. The following definitions make these properties more precise.

Definition 1. *We say that a vertex j is* exclusive *for $O_i \in M''$, and write $j \in E_i$, if we have that* $\Pr[O_{-i} \to j \mid O_i \to j \wedge S \to j \wedge H \not\to j] < 48\epsilon$.

Definition 2. *We say that a vertex j is* good *for $O_i \in M''$, and write $j \in G_i$, if we have that* $\Pr[O_i \to j \wedge S \to j \wedge H \not\to j] > \frac{1}{100}\Pr[O_i \to j \wedge H \not\to j]$.

The following lemma, proved in Appendix B, states that a large fraction of the influence of M'' is generated by events in which a node j is influenced by an O_i for which it is exclusive and good.

Lemma 7. *There exists const. c_0 s.t.* $\sum_i \sum_{j \in G_i \cap E_i} \Pr[O_i \to j \wedge S \to j \wedge H \not\to j] > c_0 \cdot \text{val(OPT)}$.

We now have all the tools needed to complete the proof of our main result.

Proof of Theorem 1: Suppose that the claim of Theorem 1 is not true, so that for any $c > 0$ there exists an input instance such that $\text{val(GRD)} \leq (1 - 1/e + c)\text{val(OPT)}$. Then by Lemma 1, Lemma 2, and Lemma 3, we can decompose OPT into L, M', and H as in the beginning of this section, and define M'' as in Lemma 5. Define S as in the beginning of the section. Choose ϵ to be arbitrarily small and write $\delta = 14\sqrt{\epsilon}$. We now claim that, for all i and j, we have

$$\Pr[O_i \to j \mid S \to j] \cdot \Pr[S \to j \mid O_i \to j] \leq 4\Pr[O_i \to S] \leq 4\delta. \tag{1}$$

The first inequality of (1) follows by considering the graph in which set S is contracted into a single vertex, then applying Lemma 6 with $y = O_i$, $z = S$, and $x = j$. The second inequality of (1) is Lemma 5.

For all i and $j \in G_i$, we know $\Pr[O_i \to j \wedge S \to j \wedge H \not\to j] \geq \frac{1}{100}\Pr[O_i \to j \wedge H \not\to j]$ from the definition of G_i, which implies $\Pr[S \to j \mid O_i \to j \wedge H \not\to j] \geq \frac{1}{100}$ and hence $\Pr[S \to j \mid O_i \to j] \geq \frac{1}{100}$ by positive correlation of connection events. Substituting into (1) we conclude that, for all i and $j \in G_i$,

$$\Pr[O_i \to j \mid S \to j \wedge H \not\to j] \leq \Pr[O_i \to j \mid S \to j] \leq 400\delta. \tag{2}$$

For all i and $j \in E_i$, we have $\Pr[O_{-i} \to j \mid O_i \to j \wedge S \to j \wedge H \not\to j] < 48\epsilon$ from the definition of E_i, which implies (by positive correlation of connectivity) that

$$\Pr[O_{-i} \to j \mid S \to j \wedge H \not\to j] < 48\epsilon. \tag{3}$$

Since $\Pr[O_{-i} \to j \mid S \to j \wedge H \not\to j] = 1 - \prod_{k \neq i}(1 - \Pr[O_k \to j \mid S \to j \wedge H \not\to j])$, a convexity argument implies that

$$\sum_{k \neq i} \Pr[O_k \to j \mid S \to j \wedge H \not\to j] < k(1 - (1 - 48\epsilon)^{1/k}) < 100\epsilon \tag{4}$$

for sufficiently small ϵ, where the last inequality holds via the Binomial approximation. Applying the union bound to (2) and (4), we conclude that $\sum_{i\colon j \in G_i \cap E_i} \Pr[O_i \to j \mid S \to j \wedge H \not\to j] < 100(\epsilon + 4\delta)$ for all j, and hence

$$\sum_{i\colon j \in G_i \cap E_i} \Pr[S \to j \wedge O_i \to j \wedge H \not\to j] \leq \sum_{i\colon j \in G_i \cap E_i} \Pr[S \to j] \cdot \Pr[O_i \to j \mid S \to j \wedge H \not\to j]$$
$$< 100(\epsilon + 4\delta)\Pr[S \to j].$$

Taking a sum over all j, we conclude

$$\sum_j \sum_{i\colon j \in G_i \cap E_i} \Pr[S \to j \wedge O_i \to j \wedge H \not\to j] < \sum_j O(\delta) \cdot \Pr[S \to j] = O(\delta) \cdot \mathrm{val}(S).$$

However, Lemma 7 implies $\sum_i \sum_{j \in G_i \cap E_i} \Pr[O_i \to j \wedge S \to j \wedge H \not\to j] > c_0 \cdot \mathrm{val}(\mathrm{OPT})$ for some constant c_0. We therefore conclude that $\mathrm{val}(S) \geq c_2 \cdot \frac{1}{\delta}\mathrm{val}(\mathrm{OPT})$ for some constant c_2, which contradicts $\mathrm{val}(S) \leq \mathrm{val}(\mathrm{OPT})$ when δ is sufficiently small. We have therefore reached contradiction, and hence we must have that $\mathrm{val}(\mathrm{GRD}) > (1 - 1/e + c)\mathrm{val}(\mathrm{OPT})$ for some constant $c > 0$. □

4 Proof of the Connectivity Lemma

In this section we prove Lemma 5, which states roughly that if many elements of the optimal solution have low probability of being connected to each other, then many of them must have low probability of being connected to the first $k/4$ elements of the greedy solution.

We first recall the formal statement of the Lemma. We let $S = \{g_1, g_2, ..., g_{k/4}\}$ be the set of the first $k/4$ vertices chosen by the greedy algorithm. We also let L, M, M', H be a decomposition of OPT satisfying the properties of Lemma 2 and Lemma 3. Lemma 5 states that there exists a set $M'' \subseteq M'$ of size at least $k/3$ such that, for each $O_i \in M''$, $\Pr[O_i \to S] < 14\sqrt{\epsilon}$.

The following lemma will be the workhorse for proving Lemma 5.

Lemma 8. *For any constant $\gamma > 0$, suppose $T = \{O_1, O_2, ..., O_{k/2}\} \subset \mathrm{OPT}$ is an arbitrary set of $k/2$ vertices chosen by the optimal algorithm such that every $O_j \in T$ satisfies $\Pr[O_j \to S] \geq \gamma$. Then there exists a set $T' \subseteq T$ of size at least $k/16$ such that for each $O_j \in T'$, $\Pr[O_j \to T \setminus \{O_j\}] \geq \gamma^2/36$.*

Proof. Consider the weighted bipartite graph H on nodes in $S \cup T$ such that for each $g_i \in S$ and $O_j \in T$ there is an edge (g_i, O_j) of weight $w_{ij} = \Pr[g_i \to O_j]$. Assume w.l.o.g. that total weight of edges incident on any node O_j is exactly γ. We say that a node $g_i \in S$ is *heavy* if the total weight incident on g_i in H is at least $\gamma/3$. Let $S_1 \subseteq S$ be the set of heavy nodes in S. We say that a node $O_j \in T$ is *heavy* if its weighted degree to nodes in S_1 is at least $\gamma/3$. Let $T_1 \subseteq T$ be the set of heavy nodes in T. A node in S or T that is not heavy is referred to as a *light* node. Note that the number of light nodes in T, say β, satisfies the following relation: $\beta(2\gamma/3) \leq (k/4)(\gamma/3)$. Thus $\beta \leq k/8$ and $|T_1| \geq k/2 - k/8 = 3k/8$. Let T_2 be the set of nodes in T_1 that that have an edge e to a node in S_1 such that the weight of e is greater than $\gamma/6$. That is, each node in T_2 contributes a weight of more than $\gamma/6$ to a single good node in S_1. Let $T_3 = T_1 \setminus T_2$. We consider separately the case in which $|T_2|$ is large and the case in which $|T_2|$ is small.

If $|T_2| \geq 5k/16$, then at least $|T_2| - |S| \geq k/16$ nodes in T_2 contribute at least $\gamma/6$ amount of weight incident on them to a node in $g \in S_1$, such that g has a total weighted degree of at least $\gamma/6$ from other nodes in T. For any of these $k/16$ nodes O_j, we have that O_j is connected to the corresponding node g with probability at least $\gamma/6$, which is then connected to another node in T with probability at least $\gamma/6$. This then implies $\Pr[O_j \to T \setminus \{O_j\}] \geq \gamma^2/36$ as required.

Otherwise, if $|T_2| < 5k/16$, then $|T_3| \geq k/16$. For each node $O_j \in T_3$, at least $\gamma/6$ of its incident weight is on edges to good nodes in S_1 that themselves receive at least $\gamma/6$ of weight from other nodes in T. For any of these $k/16$ nodes $O_j \in T_3$, we have that O_j is connected to one of these corresponding nodes in S_1, say g, with probability at least $\gamma/6$, which is then connected to another node in T with probability at least $\gamma/6$. This implies $\Pr[O_j \to T \setminus \{O_j\}] \geq \gamma^2/36$ as required. $\square$

Suppose there exists some $T \subset M'$ with $|T| = k/2$ such that $\Pr[O_j \to S] \geq 14\sqrt{\epsilon}$ for each $O_j \in T$. Then by Lemma 8 there exists some $O_j \in T$ such that $\Pr[O_j \to T \backslash \{O_j\}] \geq (14\sqrt{\epsilon})^2/36 > 5\epsilon$. But then this implies $\Pr[O_j \to O_{-j}] > 5\epsilon$, which contradicts the definition of M' (by Lemma 3). We conclude that there cannot exist such a set T. Since $|M'| < \frac{5}{6}k$, this implies that there exist at least $k/3$ nodes $O_i \in M'$ such that $\Pr[O_i \to S] < 14\sqrt{\epsilon}$, completing the proof of Lemma 5.

5 Proof of the XYZ Lemma

In this section we prove the XYZ Lemma, which was stated earlier as Lemma 6. We'll begin by restating Lemma 6.

Lemma 6 (XYZ Lemma). *In any undirected graph G with a probability function on the edges, for any 3 vertices x, y, and z, we have* $\Pr[y \to z] \geq \frac{1}{4} \cdot \Pr[x \to z \mid x \to y] \cdot \Pr[x \to y \mid x \to z]$.

We note that the inequality in the XYZ Lemma is tight up to the factor of 4. That is, there are instances in which $\Pr[y \to z] \leq \Pr[x \to z \mid x \to y] \cdot \Pr[x \to y \mid x \to z]$, even when the connection probabilities are bounded away from 1. To see this, consider a line graph on three nodes, with endpoints y and z each connected to a middle vertex x. In this example the events $x \to z$ and $x \to y$ are independent, and the intersection of those two events is precisely the event $y \to z$. Thus, for this example, we have $\Pr[y \to z] = \Pr[x \to z] \cdot \Pr[x \to y] = \Pr[x \to z \mid x \to y] \cdot \Pr[x \to y \mid x \to z]$.

We begin our proof of the XYZ Lemma by defining some new notation and making simple observations. We then provide a high-level description of the proof. The technical part of the argument is a sequence of claims that bound the probability of various connection events. We will then complete the proof by relating the event that y and z are connected to an alternative sequence of connection events.

Notation and Observations. Viewing G as a probability distribution over graphs, we say a *realization* of G is a subgraph of G in which each edge is present independently with probability as specified in G. We write $H \sim G$ to mean that H is drawn from the distribution G. Given a realization H and subsets of vertices A and B, write $A \to_H B$ for the event that some vertex in A is connected to some vertex in B. We will often abuse notation and represent a singleton set $\{a\}$ by simply a, so that if a and b are vertices then $a \to_H b$ is the event that a is connected to b. We also extend this notation to more than two parameters, so that if a, b, c are vertices of H then $a \to_H b \to_H c$ is the event that all three vertices are connected in H, etc.

We next define an ordering σ over subgraphs of G that will be used throughout the proof. Fix an arbitrary ordering over the edges of G. We then order the subgraphs of G lexicographically according to this edge order. We write $H_1 <_\sigma H_2$ to mean that H_1 occurs before H_2 in this ordering. Also, we say H is the *σ-minimal subgraph* satisfying property P if H satisfies P and $H \leq_\sigma H'$ for any H' satisfying P. Finally, we will write $\top$ for a null entry in the ordering σ, with $\top >_\sigma H$ for any subgraph H.

Given vertices a and b, and a realization H of G, we write $\textsc{Min}_H(a, b)$ for the σ-minimal subgraph of H that connects node a to node b. That is, $\textsc{Min}_H(a, b)$ is the σ-minimal connected subgraph that contains both a and b. We will define $\textsc{Min}_H(a, b) = \top$ if no such subgraph exists. Note that $a \to_H b$ is precisely the event that $\textsc{Min}_H(a, b) \neq \top$. Note also that if H and H' are two subgraphs with $a \to_H b$, but $a \not\to_{H'} b$, then $\textsc{Min}_H(a, b) <_\sigma \textsc{Min}_{H'}(a, b)$.

We further extend the notation Min in two ways. First, we will allow more than two parameters, so that $\textsc{Min}_H(a,b,c)$ is the σ-minimal connected subgraph of H that contains a, b, and c. Also, when there are only two parameters, we allow one or both parameters to be a set of vertices. In this case, $\textsc{Min}_H(A,B)$ is the σ-minimal subgraph of H that connects some node in A to some node in B, or $\top$ if no such subgraph exists. Note that if $a \to_H b$, then $\textsc{Min}_H(a,b)$ is always a simple path from a to b. Also, if $a \to_H b \to_H c$ then $\textsc{Min}_H(a,b,c)$ is always a tree whose leaves are a subset of $\{a,b,c\}$. Finally, if $a \to_H b \to_H c$, then we always have $\textsc{Min}_H(a,b) \subseteq \textsc{Min}_H(a,b,c)$.

We will write $\Gamma(a,b)$ for the distribution of $\textsc{Min}_H(a,b)$ (with randomness over $H \sim G$), conditioned on $a \to_H b$. Note that $\emptyset$ has probability 0 under $\Gamma(a,b)$, except for the trivial case $a = b$. Write $\mathrm{Supp}(a,b)$ for the support of $\Gamma(a,b)$. We extend the definition of Γ and Supp to allow three parameters that are all singletons, or two parameters that are sets of vertices, in the same way as Min. We will also write $\Gamma(a,b|c)$ for the distribution of $\textsc{Min}_H(a,b)$ when we condition on the event that a, b, and c are all in the same connected component of H. Equivalently, $\Gamma(a,b|c)$ is the distribution over subgraphs defined by first drawing γ from $\Gamma(a,b,c)$, then considering the minimal subgraph of γ' that connects vertex a to vertex b.

In our proof, we will be interested primarily in the nature of subgraph $\textsc{Min}_H(x,y,z)$. We will sometimes abuse notation slightly and think of $\textsc{Min}_H(x,y,z)$ as a set of edges, rather than a subgraph. Note that if $x \to_H y \to_H z$ then $\textsc{Min}_H(x,y,z) = \textsc{Min}_H(x,y) \cup \textsc{Min}_H(z, \textsc{Min}_H(x,y))$. Furthermore, $\textsc{Min}_H(z, \textsc{Min}_H(x,y))$ is a simple path from z to some vertex $j \in \textsc{Min}_H(x,y)$. Given some $\gamma \in \mathrm{Supp}(x,y,z)$, we will write $J(\gamma)$ for this vertex j. In other words, if $\gamma = \textsc{Min}_H(x,y,z)$ and $j = J(\gamma)$, then $\gamma = \textsc{Min}_H(x,j) \cup \textsc{Min}_H(y,j) \cup \textsc{Min}_H(z,j)$.

Proof Overview. Before going into the details of the proof, let us describe our high-level approach. We will bound the probability that y and z are in the same component in a realization of G by considering a sequence of events defined via a thought experiment. We imagine drawing $\hat{H}$ as a realization of G, and then (separately and independently) drawing some $\gamma_{xy} \sim \Gamma_{xy}$. We emphasize that γ_{xy} is unrelated to the graph $\hat{H}$; in particular, it is not necessarily the case that $\textsc{Min}_{\hat{H}}(x,y) = \gamma_{xy}$. We will first consider the event that $\hat{H}$ contains a path μ from z to a node in γ_{xy}. We'll show that this event occurs with probability at least $\Pr_{H\sim G}[x \to_H z \mid x \to_H y]$. Assuming that this occurs, we will then consider the path from x to z consisting of μ plus part of γ_{xy}; call this path γ'_{xz}. Again, this path does not necessarily exist in $\hat{H}$. We then consider the event that $\hat{H}$ contains a path μ' from y to a node in γ'_{xz}. We will show that this event occurs with probability at least $\frac{1}{2}\Pr_{H\sim G}[x \to_H y \mid x \to_H z]$. Finally, we consider the probability that paths μ and μ' intersect; a symmetry argument will establish that this occurs with probability at least $1/2$. Combining these probability bounds will prove the lemma, since these events imply that $\mu \cup \mu'$ is a connected subgraph of $\hat{H}$ that contains a path from y to z. An important subtlety in the proof is that we must choose the paths μ and μ' carefully in order to apply our desired symmetry argument. We will therefore study slightly modified versions of the first two events, crafted to ensure that the distributions over paths μ and μ' are nicely behaved.

Technical Probabilistic Bounds. We begin by establishing some bounds on the probability of certain connection events. Our first claim establishes that the probability that $z \to_H x$, given $x \to_H y$, is the same as the probability of the following event. Draw a random path γ from the distribution $\Gamma(x,y)$ of minimal $x - y$ paths, then draw $H \sim G$ conditioned on H not containing an $x - y$ path lexicographically less than γ; the event is that z is connected to γ in this graph H.

Claim 1.

$$\Pr_{\substack{\gamma\sim\Gamma(x,y),\\ H\sim G\colon\ \textsc{Min}_H(x,y)\geq_\sigma \gamma}}[z \to_H \gamma] = \Pr_H[z \to_H x \mid x \to_H y].$$

Proof.

$$\begin{aligned}
\Pr_H[z \to_H x \mid x \to_H y] &= \sum_{\gamma \in \mathrm{Supp}(x,y)} \Pr_H[\textsc{Min}_H(x,y) = \gamma \mid x \to_H y] \cdot \Pr_H[z \to_H \gamma \mid \textsc{Min}_H(x,y) = \gamma] \\
&= \Pr_{\substack{\gamma \sim \Gamma(x,y), \\ H \sim G\colon\ \textsc{Min}_H(x,y) = \gamma}} [z \to_H \gamma] \\
&= \Pr_{\substack{\gamma \sim \Gamma(x,y), \\ H \sim G\colon\ \textsc{Min}_H(x,y) \geq_\sigma \gamma}} [z \to_H \gamma]
\end{aligned}$$

where the first equality is simply expanding event $x \to_H y$ by conditioning on the identity of $\textsc{Min}_H(x,y)$, the second equality is the definition of $\Gamma(x,y)$, and the last equality follows because the existence of a path from z to γ does not depend on whether or not the edges in γ are actually present in realization H, only on the fact that no lexicographically lesser path from x to y exists in H. □

We next show that if we modify Claim 1 so that path γ is drawn from the distribution $\Gamma(x,y|z)$ rather than $\Gamma(x,y)$, then this can only increase the probability that z is connected to γ. Intuitively, this is because $\Gamma(x,y|z)$ favors paths that are more likely to be connected to z, as it draws from the distribution of subgraphs connecting all of x, y, z restricted to paths from x to y.

Claim 2.

$$\Pr_{\substack{\gamma \sim \Gamma(x,y|z), \\ H \sim G\colon\ \textsc{Min}_H(x,y) \geq_\sigma \gamma}} [x \to_H \gamma] \geq \Pr_H[z \to_H x \mid x \to_H y].$$

Proof. We have the following, which follows precisely as in Claim 1 except for the inequality, which is proven below.

$$\begin{aligned}
\Pr_H[z \to_H x \mid x \to_H y] &= \sum_{\gamma \in \mathrm{Supp}(x,y)} \Pr_H[\textsc{Min}_H(x,y) = \gamma \mid x \to_H y] \cdot \Pr_H[z \to_H \gamma \mid \textsc{Min}_H(x,y) = \gamma] \\
&= \Pr_{\substack{\gamma \sim \Gamma(x,y), \\ H \sim G\colon\ \textsc{Min}_H(x,y) = \gamma}} [z \to_H \gamma] \\
&\leq \Pr_{\substack{\gamma \sim \Gamma(x,y|z), \\ H \sim G\colon\ \textsc{Min}_H(x,y) = \gamma}} [z \to_H \gamma] \\
&= \Pr_{\substack{\gamma \sim \Gamma(x,y|z), \\ H \sim G\colon\ \textsc{Min}_H(x,y) \geq_\sigma \gamma}} [z \to_H \gamma].
\end{aligned}$$

As in Claim 1, the last equality follows because the identity of the minimal path from z to γ does not depend on whether or not the edges in γ are actually present, only on the fact that no lexicographically lesser path from x to y exists in H.

We now prove the inequality. For notational convenience, let

$$p_{xy}(\gamma) := \Pr_{H \sim G}[z \to_H \gamma \mid \textsc{Min}_H(x,y) = \gamma],$$

let $\Gamma_{xy}(\gamma)$ be the probability of γ under $\Gamma(x,y)$, and let $\Gamma_{xy|z}(\gamma)$ be the probability of γ under $\Gamma(x,y|z)$. Under this notation, the inequality to prove is

$$\sum_\gamma \Gamma_{xy}(\gamma) \cdot p_{xy}(\gamma) \leq \sum_\gamma \Gamma_{xy|z}(\gamma) \cdot p_{xy}(\gamma)$$

To see this, note that $\Gamma_{xy|z}(\gamma) = \frac{\Gamma_{xy}(\gamma) p_{xy}(\gamma)}{\sum_{\gamma'} \Gamma_{xy}(\gamma') p_{xy}(\gamma')}$. It therefore suffices to show that

$$\left[\sum_\gamma \Gamma_{xy}(\gamma) p_{xy}(\gamma) \right]^2 \leq \sum_\gamma \Gamma_{xy}(\gamma) p_{xy}(\gamma)^2.$$

This follows from Cauchy-Schwarz, since

$$\begin{aligned}\sum_{\gamma}\Gamma_{xy}(\gamma)p_{xy}(\gamma) &= \sum_{\gamma}\sqrt{\Gamma_{xy}(\gamma)}\left(\sqrt{\Gamma_{xy}(\gamma)}p_{xy}(\gamma)\right)\\ &\leq \left(\sqrt{\sum_{\gamma}\Gamma_{xy}(\gamma)}\right)\left(\sqrt{\sum_{\gamma}\Gamma_{xy}(\gamma)p_{xy}(\gamma)^2}\right)\\ &= \sqrt{\sum_{\gamma}\Gamma_{xy}(\gamma)p_{xy}(\gamma)^2},\end{aligned}$$

where we used the fact that $\sum_{\gamma}\Gamma_{xy}(\gamma) \leq 1$. □

Our final claim makes two modifications to Claim 2. First, we swap the roles of z and y for notational convenience that will become apparent later; this does not affect the argument[2]. Second, we modify Claim 2 so that, after drawing path γ from $\Gamma(x, z|y)$, we also draw a path μ from y to γ, according to distribution $\Gamma(y, \gamma)$, and reveal that H contains no path from y to γ lexicographically less than μ. Our claim is that, in expectation, this will reduce the probability that y is connected to γ by at most $\frac{1}{2}$.

More formally, let $\mathcal{E}$ be the following event. Draw $\gamma' \sim \Gamma(x, z|y)$, then draw $\mu \sim \Gamma(y, \gamma)$. Draw $H \sim G$ subject to $\textsc{Min}_H(x, z) \geq_\sigma \gamma'$ and $\textsc{Min}_H(y, \gamma') \geq_\sigma \mu$. Then $\mathcal{E}$ is the event that $y \to_H \gamma'$ under this distribution of H.

Claim 3. *For $\mathcal{E}$ as defined above,*

$$\Pr[\mathcal{E}] \geq \frac{1}{2}\Pr_H[y \to_H x \mid x \to_H z].$$

Proof. Let $\mathcal{E}'$ be defined similarly to $\mathcal{E}$, except that when we draw H we do not condition on $\textsc{Min}_H(y, \gamma') \geq_\sigma \mu$; only that $\textsc{Min}_H(x, z) \geq_\sigma \gamma'$. Then Claim 2 (swapping the roles of y and z) precisely states that $\Pr[\mathcal{E}'] \geq \Pr_H[y \to_H x \mid x \to_H z]$. So it suffices to show that $\Pr[\mathcal{E}] \geq \frac{1}{2}\Pr[\mathcal{E}']$.

For fixed γ', let $D(\gamma')$ be the distribution over H conditioned on $\textsc{Min}_H(x, z) \geq_\sigma \gamma'$. From the definition of $\mathcal{E}$, we have

$$\begin{aligned}\Pr[\mathcal{E}] &= \mathrm{E}_{\gamma'\sim\Gamma(x,z|y),\ \mu\sim\Gamma(y,\gamma')}\left[\Pr_{H\sim D(\gamma')}[y \to_H \gamma' \mid \textsc{Min}_H(y, \gamma') \geq_\sigma \mu]\right]\\ &= \mathrm{E}_{\gamma'\sim\Gamma(x,z|y),\ \mu\sim\Gamma(y,\gamma')}\left[\frac{\Pr_{H\sim D(\gamma')}[y \to_H \gamma' \wedge \textsc{Min}_H(y, \gamma') \geq_\sigma \mu]}{\Pr_{H\sim D(\gamma')}[\textsc{Min}_H(y, \gamma') \geq_\sigma \mu]}\right]\\ &\geq \mathrm{E}_{\gamma'\sim\Gamma(x,z|y),\ \mu\sim\Gamma(y,\gamma')}\left[\Pr_{H\sim D(\gamma')}[y \to_H \gamma' \wedge \textsc{Min}_H(y, \gamma') \geq_\sigma \mu]\right]\\ &= \mathrm{E}_{\gamma'\sim\Gamma(x,z|y),\ \mu\sim\Gamma(y,\gamma')}\left[\Pr_{H\sim D(\gamma')}[\textsc{Min}_H(y, \gamma') \geq_\sigma \mu \mid y \to_H \gamma'] \cdot \Pr_{H\sim D(\gamma')}[y \to_H \gamma']\right]\\ &\geq \mathrm{E}_{\gamma'\sim\Gamma(x,z|y),\ \mu\sim\Gamma(y,\gamma')}\left[\frac{1}{2}\Pr_{H\sim D(\gamma')}[y \to_H \gamma']\right]\\ &= \frac{1}{2}\Pr[\mathcal{E}']\end{aligned}$$

as required, where the final inequality follows by symmetry: $\textsc{Min}_H(y, \gamma')$ and μ are drawn from the same distribution given γ', and hence the probability that $\textsc{Min}_H(y, \gamma') \geq_\sigma \mu$ is at least $\frac{1}{2}$. □

Let $\mathcal{E}$ be the event from Claim 3, and let γ' and μ' be defined as in event $\mathcal{E}$. Let $\gamma = \gamma' \cup \mu$; note that $\gamma \in \mathrm{Supp}(x, y, z)$. Let $j = J(\gamma)$. Conditioning on event $\mathcal{E}$, let $\mu' = \textsc{Min}_H(y, \gamma')$. Recall

[2] We stated Claim 2 without this change to more easily draw a parallel between its proof and the proof of Claim 1.

that γ' is a simple path from x to z that contains vertex $J(\gamma)$. Let π_z denote the probability (over all randomness in event $\mathcal{E}$, but conditioning on $\mathcal{E}$) that μ' intersects γ' on the subpath connecting $J(\gamma)$ to z. That is, μ' intersects γ' no "farther" from z than μ does. Then note that we can assume without loss of generality that $\pi_z \geq \frac{1}{2}$; if not, we simply relabel vertices x and z. Note that this is the only point at which we consider relabeling the vertices x, y, z.

Assumption: $\pi_z \geq \frac{1}{2}$.

We now have the tools we need to complete the proof of Lemma 6.

Proof of Lemma 6. Consider the following thought experiment. Draw $\hat{H}$ as a realization of G, and then separately draw $\gamma_{xy} \sim \Gamma(x, y)$. Let $\mathcal{E}_1$ be the event $z \rightarrow_{\hat{H}} \gamma_{xy}$. We will also define a refined version of event $\mathcal{E}_1$, which we call $\mathcal{E}_1'$. This event $\mathcal{E}_1'$ will imply $\mathcal{E}_1$, but will also satisfy the stronger property that, conditioning on event $\mathcal{E}_1'$, the connection between z and γ_{xy} occurs via a path μ such that $\gamma_{xy} \cup \mu$ is distributed according to $\Gamma(x, y, z)$. Roughly speaking, $\mathcal{E}_1'$ is the event that z is connected to γ_{xy} even after certain edges have been removed from $\hat{H}$.

We now define $\mathcal{E}_1'$ more formally, as follows. Given γ_{xy}, we will define a randomized mapping Ψ on the set of subgraphs of G that include γ_{xy}, with the following properties:

1. $\Psi(H) \subseteq H$ for all H,
2. $\textsc{Min}_{\Psi(H)}(x, y) = \gamma_{xy}$ for all H, and
3. the distribution over $\Psi(H)$, with randomness taken over $H \sim G$ (conditional on H containing γ_{xy}) and in Ψ, is precisely the uniform distribution over the set of subgraphs H' for which $\textsc{Min}_{H'}(x, y) = \gamma_{xy}$.

Before proving the existence of Ψ, let us finish the definition of $\mathcal{E}_1'$. Write H' for the random variable representing $\Psi(\hat{H} \cup \gamma_{xy})$. The event $\mathcal{E}_1'$ will then be the event that $z \rightarrow_{H'} \gamma_{xy}$, with randomness taken over $\hat{H}$, γ_{xy}, and Ψ. Note that by property 1 of Ψ, $\mathcal{E}_1'$ implies $\mathcal{E}_1$. Moreover, property 3 implies that $\Psi(\hat{H} \cup \gamma_{xy})$ is distributed precisely as a realization drawn from G, conditional on γ_{xy} being the minimal path connecting x to y.

To see that an appropriate mapping Ψ exists, note that condition 3 states that Ψ should map the uniform distribution over graphs including γ_{xy} to the uniform distribution over graphs H for which $\textsc{Min}_H(x, y) = \gamma_{xy}$. The former distribution stochastically dominates the latter, since the difference in the supports of the distributions is upward-closed (with respect to the addition of edges). It is therefore possible to map the former distribution to the latter by shifting probability mass from subgraphs H to subgraphs H' with $H' \subseteq H$ [22]. Such a transformation precisely defines a mapping Ψ satisfying the required properties.

Claim 4. $\Pr[\mathcal{E}_1'] = \Pr_H[z \rightarrow_H x \mid x \rightarrow_H y]$.

Proof. From the definition of $\mathcal{E}_1'$, plus the observation that (for fixed γ_{xy}, and randomness taken over $\hat{H}$) the graph $H' = \Psi(\hat{H})$ is distributed uniformly over the set of graphs H for which γ_{xy} would be the minimal path from x to y (if present), we have $\Pr[\mathcal{E}_1'] = \Pr_{\substack{\gamma \sim \Gamma(x,y), \\ H \sim G:\ \textsc{Min}_H(x,y) \geq_\sigma \gamma}}[z \rightarrow_H \gamma]$. Claim 1 then implies the desired result. □

For the remainder of the argument we will condition on the event $\mathcal{E}_1'$. Let $\mu = \textsc{Min}_{H'}(z, \gamma_{xy})$ be the minimal path from z to γ_{xy} in H'. Let $\gamma_{xyz}' = \mu \cup \gamma_{xy}$, and let $j = J(\gamma_{xyz}')$. Let γ_{yj}' be the path from y to j contained in γ_{xyz}'. Thinking of γ_{xyz}' as a random variable (over the randomness in $\hat{H}$, γ_{xy}, and H'), we have $\gamma_{xyz}' \sim \Gamma(x, y, z)$. This is because $\gamma_{xy} \sim \Gamma(x, y|z)$ (since we condition on event $\mathcal{E}_1'$) and $\mu \sim \Gamma(z, \gamma_{xy})$. Let γ'_{xz} be the path from x to z contained in γ'_{xyz}. Again thinking of $\Gamma(x, z|y)$ as a random variable over the randomness in γ_{xy}, $\hat{H}$, and H', we have that $\gamma'_{xz} \sim \Gamma(x, z|y)$.

Let $\mathcal{E}_2$ be the event $y \to_{H'} \gamma'_{xz}$.

Claim 5. $\Pr[\mathcal{E}_2 \mid \mathcal{E}_1'] \geq \frac{1}{2}\Pr_H[y \to_H x \mid x \to_H z]$

Proof. Condition on event $\mathcal{E}_1'$, and on the identity of γ'_{xyz} from the discussion immediately preceeding this claim. Think now of H' as a random variable, drawn from G subject to the conditions imposed by $\mathcal{E}_1'$ and the identity of γ'_{xyz}. These conditions are precisely that H' contains μ (the path from z to $J(\gamma'_{xyz})$ in γ'_{xyz}) and that $\textsc{Min}_{H'}(x, y, z) \geq_\sigma \gamma'_{xyz}$. That is, H' does not contain a subgraph lexicographically less than γ'_{xyz} that connects x, y, and z.

The condition $\textsc{Min}_{H'}(x, y, z) \geq_\sigma \gamma'_{xyz}$ is precisely equivalent to the following pair of conditions: $\textsc{Min}_{H'}(x, z) \geq_\sigma \gamma'_{xz}$ and $\textsc{Min}_{H'}(y, \gamma'_{xz}) \geq_\sigma \gamma'_{yj}$. To see this, note that if H' contains a lexicographically smaller sugraph in $\mathrm{Supp}(x, y, z)$, then this occurs either because it contains a lesser path from x to z than γ'_{xz}, or a lesser path from y to γ'_{xz} than γ'_{yj}.

But we can now apply Claim 3, since this distribution of H' is precisely the distribution described in event $\mathcal{E}$. We therefore have $\Pr[\mathcal{E}_2] = \Pr[\mathcal{E}] \geq \frac{1}{2}\Pr_H[y \to_H x \mid x \to_H z]$, as required. □

Conditioned on $\mathcal{E}_2$ occurring, there is a minimal path μ' from y to γ'_{xz} in H', call it μ'. Let $\mathcal{E}_3$ be the event that μ and μ' intersect. Note that $\mathcal{E}_3$ certainly occurs if μ' intersects γ'_{xz} on the subpath of γ'_{xz} connecting z to j, since that subpath is precisely μ. But this probability is precisely what we defined to be π_z, which is at least $\frac{1}{2}$ by assumption. We therefore have $\Pr[\mathcal{E}_3 \mid \mathcal{E}_1' \wedge \mathcal{E}_2] \geq \frac{1}{2}$.

Observe that $\mathcal{E}_1' \cap \mathcal{E}_2 \cap \mathcal{E}_3$ imply that y and z are in the same component in H. We therefore have

$$\begin{aligned}\Pr_H[y \to_H z] &\geq \Pr[\mathcal{E}_1' \cap \mathcal{E}_2 \cap \mathcal{E}_3] \\ &= \Pr[\mathcal{E}_1'] \cdot \Pr[\mathcal{E}_2 \mid \mathcal{E}_1'] \cdot \Pr[\mathcal{E}_3 \mid \mathcal{E}_1' \wedge \mathcal{E}_2] \\ &\geq \frac{1}{4}\Pr[x \to_H z \mid x \to_H y]\Pr[x \to_H y \mid x \to_H z],\end{aligned}$$

using Claims 4 and 5. This completes the proof of the Lemma 6.

6 APX-Hardness of Undirected Influence Maximization

We now establish that the problem of undirected influence maximization is APX-hard.

Theorem 2. *There exists an $\epsilon_0 > 0$ such that it is NP-hard to obtain a $(1 - \epsilon_0)$-approximation to the undirected influence maximization problem.*

Proof. We reduce from the problem of finding a vertex cover in cubic graphs (all vertices have degree exactly 3). It is known that there exists an $\epsilon > 0$ and a parameter $K(n)$ so that it is NP hard to decide whether an n-vertex cubic graph has a vertex cover of size $\leq K(n)$ (a Yes-instance) or if any subset of $K(n)$ vertices leaves at least ϵm edges uncovered (a No-instance) [1].

Given a vertex cover instance $\langle G = (V, E), K(n)\rangle$ on n vertices and m edges, we construct an instance $\langle G'(V', E'), U, p, w, k\rangle$ of undirected influence maximization as follows. The graph $G'(V', E')$ is an undirected graph where (a) the vertex set $V' = L \cup R$ contains a vertex v for each vertex $v \in V$, as well as a vertex e for each edge $e \in E$, and (b) the edge set E' contains an edge (v, e) iff e is incident on vertex v in the graph $G(V, E)$. Thus G' is a bipartite graph where the left partition L corresponds to vertices in G and the right partition R corresponds to edges in G, and each vertex in L has degree 3 while each vertex in R has degree 2. Finally, the set U of allowed seed vertices is precisely the set L, the cascade probability function p is defined to be uniform i.e. $p(e) = p$ for some fixed p for every edge $e \in E'$, the weight function w is defined to be 0 for vertices in L and 1 for vertices in R, and $k = K(n)$. We will argue that for a suitable choice of the parameter p, there exists constants $0 < \epsilon_1 < \epsilon_2$ such that whenever G is a Yes-instance of vertex cover, then the optimal value for the influence maximization problem is at least $(1 - \epsilon_1)|R|$, and whenever G is a No-instance of vertex cover, then the optimal value for the influence maximization problem is at most $(1 - \epsilon_2)|R|$. It follows that the undirected influence maximization problem is APX-hard.

To analyze the optimal value in the instance $G'(V', E')$, we analyze three quantities of interest for each edge e. For any set $S \subseteq L$ of seed vertices, and a vertex $e = (u, v) \in R$, we define (a) ϕ_0 as the probability that e is activated when neither u nor v are in S, (b) ϕ_1 as the probability that edge e is activated when only one of u and v is in S, and (c) ϕ_2 as the probability that edge e is activated when both u and v are in S. Then it is easy to verify that

$$\begin{aligned}
\phi_0 &\leq p\left(p^2 + (1-p^2)p^2\right) + \left(1 - p\left(p^2 + (1-p^2)p^2\right)\right)\left(p\left(p^2 + (1-p^2)p^2\right)\right) \\
\phi_1 &\leq p + (1-p)\left(p^2 + (1-p^2)p^2\right) \\
\phi_2 &= p + (1-p)p,
\end{aligned}$$

where the bound on $\phi_1(e)$ is exact whenever the seed set S corresponds to a vertex cover in G. We are now ready to bound the optimal value based for influence maximization on whether or not G was a Yes-instance of vertex cover. Let $k = K(n) = m/3(1+\alpha)$ for some $\alpha \in [0, 1]$ (since any vertex cover in a cubic graph G must have size at least $m/3$ and at most $n \leq 2m/3$).

Yes-instance analysis: When $G(V, E)$ is a Yes-instance, there is a set $S \subseteq V$ of size k such that every edge in E has at least one end-point in S. We will choose the set S as our set of seed vertices in $G'(L \cup R, E')$. Clearly, every vertex $e \in R$ is adjacent to at least one vertex in our seed set S. As degree of each vertex is exactly 3, it must be the case that $3(k - m/3) = \alpha m$ edge vertices in R are adjacent to two vertices in S. Thus $\mathrm{val}(S) \geq (1-\alpha)m\phi_1 + \alpha m\phi_2$. Let $V_{\mathsf{yes}} = (1-\alpha)m\phi_1 + \alpha m\phi_2$ denote the above lower bound on the optimal solution value when $G(V, E)$ is a Yes-instance.

No-instance analysis: When $G(V, E)$ is a No-instance, for any set $S \subseteq V$ of size k, at least ϵm edges are left uncovered. It then follows that for any set $S \subseteq L$ in G', at least ϵm vertices in R have no chosen seed vertex adjacent to them. If for a set S, the number of uncovered edges is exactly $\epsilon' m$ for some $\epsilon' \geq \epsilon$, then $(1-\alpha-2\epsilon')$-fraction of edges are covered exactly once and $(\alpha+\epsilon')$-fraction of edges are covered twice. Thus for such a set S of seed vertices, we have $\mathrm{val}(S) \leq \epsilon' m\phi_0 + (1-\alpha-2\epsilon')m\phi_1 + (\alpha+\epsilon')m\phi_2$. Note that $\phi_2 - \phi_1 \leq \phi_1 - \phi_0$ for all $p \in [0, 1]$ since $(\phi_1 - \phi_0) - (\phi_2 - \phi_1) = p^2(1-p^2)^4$ which is non-negative for all $p \in [0, 1]$. It follows that the preceding upper bound on $\mathrm{val}(S)$ is maximized when $\epsilon' = \epsilon$. Hence, for any set S, we have $\mathrm{val}(S) \leq \epsilon m\phi_0 + (1-\alpha-2\epsilon)m\phi_1 + (\alpha+\epsilon)m\phi_2$. Let V_{no} denote this lower bound on the optimal solution value when $G(V, E)$ is a No-instance.

Now let edge cascade probability $p = 1 - \epsilon/D$ where ϵ is the gap parameter for the vertex cover instance and $D > 1$ is a parameter whose value will be determined later. Then a simple calculation shows that $V_{\mathsf{yes}} - V_{\mathsf{no}} \geq \left(\frac{16\epsilon^5}{D^4} - \frac{64\epsilon^6}{D^5}\right) m$. Choosing $D = 8\epsilon$, we get $V_{\mathsf{yes}} - V_{\mathsf{no}} \geq \frac{\epsilon}{8^3} m$, which (due to the known hardness of finding a vertex cover in cubic graphs) proves Theorem 2. □

References

[1] Paola Alimonti and Viggo Kann. Some apx-completeness results for cubic graphs. pages 123–134, 2000.

[2] Shishir Bharathi, David Kempe, and Mahyar Salek. Competitive influence maximization in social networks. In *WINE*, pages 306–311, 2007.

[3] Duncan S Callaway, Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Network robustness and fragility: Percolation on random graphs. *Physical review letters*, 85(25):5468, 2000.

[4] Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD*, pages 1029–1038, 2010.

[5] Charles J Colbourn. *The combinatorics of network reliability.* Oxford University Press, Inc., 1987.

[6] Peter S. Dodds and D.J.Watts. A generalized model of social and biological contagion. *Journal of Theoretical Biology*, 232(4):587–694, 2005.

[7] Pedro Domingos and Matthew Richardson. Mining the network value of customers. In *KDD*, pages 57–66, 2001.

[8] J. Goldenberg, B. Libai, and E. Mulle. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Mark. Let.*, pages 221–223, 2001.

[9] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. *TKDD*, 5(4):21, 2012.

[10] Sanjeev Goyal and Michael Kearns. Competitive contagion in networks. In *STOC*, pages 759–774, 2012.

[11] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, (83):1420–1443, 1978.

[12] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[13] David Kempe, Jon M. Kleinberg, and Éva Tardos. Influential nodes in a diffusion model for social networks. In *ICALP*, pages 1127–1138, 2005.

[14] Harry Kesten. The critical probability of bond percolation on the square lattice equals 1/2. *Communications in mathematical physics*, 74(1):41–59, 1980.

[15] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie S. Glance, and Matthew Hurst. Patterns of cascading behavior in large blog graphs. In *SDM*, 2007.

[16] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using internet chain-letter data. *PNAS*, 105(12):4633–4638, 2008.

[17] Christian D Lorenz and Robert M Ziff. Precise determination of the bond percolation thresholds and finite-size scaling corrections for the sc, fcc, and bcc lattices. *Physical Review E*, 57(1):230, 1998.

[18] Elchanan Mossel and Sebastien Roch. On the submodularity of influence in social networks. In *STOC*, pages 128–134, 2007.

[19] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

[20] Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD*, pages 61–70, 2002.

[21] E. Rogers. *Diffusion of Innovations.* Free Press, 5th edition, 2003.

[22] Lars Peter sterdal. The mass transfer approach to multivariate discrete first order stochastic dominance: Direct proof and implications. *Journal of Mathematical Economics*, 46(6):1222–1228, November 2010.

A Omitted Proofs

Proof of Lemma 1: Let $\gamma \in (0,1)$ be a constant to be determined later. We analyze the performance of the greedy algorithm by comparing the value of the first $\gamma\delta k$ sets relative to the sets in the collection X, and then the rest relative to the residual value of the optimal. The value of the first $\gamma\delta k$ sets chosen by the greedy algorithm, say a collection Y, is at least

$$\text{val}(Y) \geq \text{val}(X)\left(1 - \frac{1}{e^{\gamma}}\right) > \frac{(1+\epsilon)\cdot\delta\cdot\text{val}(\text{OPT})}{k}\left(1 - \frac{1}{e^{\gamma}}\right),$$

where the first inequality follows by applying the standard greedy analysis and measuring val(Y) relative to val(X), and the second inequality follows from our assumption about the set X.

The next $(1-\gamma\delta)k$ sets chosen by the greedy algorithm, say a collection Z, gets an additional contribution of at least

$$\text{val}(Z) \geq (\text{val}(\text{OPT}) - \text{val}(Y))\left[1 - \frac{1}{e^{1-\gamma\delta}}\right].$$

Thus total value of the sets chosen by the greedy must be at least

$$\text{val}(Y) + \text{val}(Z) \geq \text{val}(\text{OPT}) - \frac{\text{val}(\text{OPT}) - \text{val}(Y)}{e^{1-\gamma\delta}}.$$

Hence the deficit of the greedy algorithm with respect to val(OPT) can be bounded by

$$\begin{aligned}\frac{\text{val}(\text{OPT}) - \text{val}(Y)}{e^{1-\gamma\delta}} &\leq \frac{\text{val}(\text{OPT}) - (1+\epsilon)\delta\text{val}(\text{OPT})(1-\frac{1}{e^{\gamma}})}{e^{1-\gamma\delta}} \\ &\leq \frac{\text{val}(\text{OPT})}{e}\left[e^{\gamma\delta}(1-(1+\epsilon)\delta) + e^{-\gamma(1-\delta)}(1+\epsilon)\delta\right]\end{aligned}$$

Let $\Gamma = \left[e^{\gamma\delta}(1-(1+\epsilon)\delta) + e^{-\gamma(1-\delta)}(1+\epsilon)\delta\right]$. To complete the proof of the lemma, it suffices to show that by choosing $\gamma = (\delta\epsilon)/2$, we can bound Γ by $\left(1 - \frac{\delta^2\epsilon^2}{4}\right)$, that is, $\Gamma < 1$ for ϵ, δ are both positive. Using the fact that $e^x \leq 1 + x + x^2$ for $|x| < 1$, we have

$$\begin{aligned}\Gamma &\leq (1+\gamma\delta+\gamma^2\delta^2)(1-(1+\epsilon)\delta) + (1-\gamma+\gamma\delta+\gamma^2(1-\delta)^2)(1+\epsilon)\delta \\ &\leq 1+\gamma\delta-\gamma(1+\epsilon)\delta+\gamma^2(\delta^2-\delta^3(1+\epsilon)+(1-\delta)^2) \\ &\leq 1-\gamma\delta\epsilon+\gamma^2(\delta^2+(1-\delta)^2) \\ &\leq 1-\gamma\delta\epsilon+\gamma^2 \\ &\leq 1-\frac{\delta^2\epsilon^2}{4},\end{aligned}$$

where the last but one inequality follows from the fact that $(\delta^2 + (1-\delta)^2) \leq 1$ for $\delta \in (0,1)$, and the last equality follows by choosing $\gamma = (\delta\epsilon)/2$. The assertion of the lemma thus follows. □

Proof of Lemma 2: We will give an iterative procedure to construct the decomposition into sets L, M and H as above, and show that the procedure succeeds unless OPT contains a set X of vertices whose normalized influence is strictly greater than 1 and $\text{val}(X) = \Omega(\text{val}(\text{OPT}))$.

Initialize Z to contain the vertices $\{O_1, ..., O_k\}$ in OPT, and initialize $L = \emptyset$. While there exists a vertex $O_i \in Z$ such that $\text{val}(O_i | Z \setminus O_i) < \frac{(1-\epsilon)\text{val}(\text{OPT})}{k}$, do $L = L \cup \{O_i\}$ and $Z = Z \setminus \{O_i\}$. If upon termination, the set L contains more than ϵk vertices, then the set $X = \text{OPT} \setminus L$ satisfies

$$\begin{aligned}
\text{val}(X) &\geq \text{val}(\text{OPT}) - |L| \cdot \frac{(1-\epsilon)\text{val}(\text{OPT})}{k} \\
&\geq \epsilon\text{val}(\text{OPT}) + k\frac{(1-\epsilon)\text{val}(\text{OPT})}{k} - |L|\frac{(1-\epsilon)\text{val}(\text{OPT})}{k} \\
&\geq \epsilon\text{val}(\text{OPT}) + |X|\frac{(1-\epsilon)\text{val}(\text{OPT})}{k} \\
&\geq \epsilon\frac{|L|+|X|}{k}\text{val}(\text{OPT}) + |X|\frac{(1-\epsilon)\text{val}(\text{OPT})}{k} \\
&\geq \epsilon\frac{|L|}{k}\text{val}(\text{OPT}) + |X|\frac{\text{val}(\text{OPT})}{k} \\
&\geq \epsilon^2\text{val}(\text{OPT}) + |X|\frac{\text{val}(\text{OPT})}{k}.
\end{aligned}$$

This gives us the desired set X with normalized influence strictly greater than 1 and $\text{val}(X) = \Omega(\text{val}(\text{OPT}))$. Assuming $|L| < \epsilon k$, we continue with the decomposition process on the remaining set Z to identify the sets M and H. Note that in this case case, we know that $\text{val}(L) \leq \epsilon \cdot \text{val}(\text{OPT})$. Let σ be an ordering of the vertices in Z created in the following manner. We choose $O_{\sigma(1)}$ to be a vertex $O \in Z$ that maximizes $\text{val}(O)$. Then $O_{\sigma(2)}$ is chosen to be a vertex $O \in Z$ that maximizes $\text{val}(O | O_{\sigma(1)})$. In general, we choose $O_{\sigma(2)}$ is chosen to be a vertex $O \in Z$ that maximizes $\text{val}(O | O_{\sigma(1)}, \ldots, O_{\sigma(i-1)})$. Now consider the largest index j such that

$$\text{val}(O_{\sigma(j)} | O_{\sigma(1)}, \ldots, O_{\sigma(j-1)}) > \frac{(1+\epsilon)\text{val}(\text{OPT})}{k}.$$

If $\text{val}(O_{\sigma(1)} \cup O_{\sigma(2)} \cup \ldots \cup O_{\sigma(j)}) \geq \epsilon^2\text{OPT}$, then $X = \{O_{\sigma(1)}, \ldots, O_{\sigma(j)}\}$ gives us a set with normalized influence strictly greater than 1 and $\text{val}(X) = \Omega(\text{val}(\text{OPT}))$. Otherwise, we continue with our decomposition and define $H = \{O_{\sigma(1)}, \ldots, O_{\sigma(j)}\}$, and $M = Z \setminus H$. We now argue that for each $O_i \in \text{OPT}'$, we have

$$\frac{(1-\epsilon)\text{val}(\text{OPT})}{k} \leq \text{val}(O_i | O_{-i}, H) \leq \text{val}(O_i | H) \leq \frac{(1+\epsilon)\text{val}(\text{OPT})}{k}.$$

The first inequality follows from the fact that $O_i \notin L$. The second inequality follows from submodularity of the influence function. The last inequality follows because $O_i \notin H$. □

Proof of Lemma 3: Let $M_1 = \{O_i \in M \mid \Pr[O_i \to H] > 2\epsilon\}$. We will first show that $|M_1| \leq \epsilon k$. To see this, note that

$$\begin{aligned}
\mathrm{val}(H) &\geq \sum_{O_i \in M} \Pr[O_i \to H]\mathrm{val}(O_i \mid M \setminus \{O_i\}) \\
&\geq 2\epsilon \cdot \sum_{O_i \in M_1} \Pr[O_i \to H]\mathrm{val}(O_i \mid M \setminus \{O_i\}) \\
&\geq 2\epsilon \cdot \sum_{O_i \in M_1} \Pr[O_i \to H]\mathrm{val}(O_i \mid M \setminus \{O_i\}, H) \\
&\geq 2\epsilon |M_1| \frac{(1-\epsilon)\mathrm{val}(\mathrm{OPT})}{k} \\
&> \epsilon^2 \cdot \mathrm{val}(\mathrm{OPT})
\end{aligned}$$

for any $\epsilon \in (0, 1/3)$ whenever $|M_1| \geq \epsilon k$. But this contradicts our assumption that $\mathrm{val}(H) \leq \epsilon^2 \mathrm{val}(\mathrm{OPT})$.

Let $M' = M \setminus M_1$. We will now show that for each vertex $O_i \in M'$, we have $\Pr[O_i \to M' \setminus \{O_i\}]$ is at most 5ϵ. We argue this as follows:

$$\begin{aligned}
\Pr[O_i \to M' \setminus \{O_i\}] &= \Pr[O_i \to M' \setminus \{O_i\} \mid O_i \to H] \cdot \Pr[O_i \to H] \\
&\quad + \Pr[O_i \to M' \setminus \{O_i\} \mid O_i \not\to H] \cdot \Pr[O_i \not\to H] \\
&\leq \Pr[O_i \to H] + \Pr[O_i \to M' \setminus \{O_i\} \mid O_i \not\to H]
\end{aligned}$$

Thus it suffices to show that $\Pr[O_i \to M' \setminus \{O_i\} \mid O_i \not\to H] \leq 3\epsilon$ for all $O_i \in M'$. To see this, we first observe that for each $O_i \in M'$ (in fact the analysis below applies to each $O_i \in M$ and not just M'), we have

$$\sum_j \Pr[O_i \to j \wedge O_i \to O_{-i} \wedge O_i \not\to H] \leq \frac{2\epsilon \mathrm{val}(\mathrm{OPT})}{k},$$

since $\mathrm{val}(O_i | O_{-i}, H) \geq (1-\epsilon)\frac{\mathrm{val}(\mathrm{OPT})}{k}$ and $\mathrm{val}(O_i | H) \leq (1+\epsilon)\frac{\mathrm{val}(\mathrm{OPT})}{k}$. So, we have

$$\begin{aligned}
\frac{2\epsilon \mathrm{val}(\mathrm{OPT})}{k} &\geq \sum_j \Pr[O_i \to j \wedge O_i \to O_{-i} \wedge O_i \not\to H] \\
&= \sum_j \Pr[O_i \to O_{-i} \mid O_i \to j \wedge O_i \not\to H] \cdot \Pr[O_i \to j \wedge O_i \not\to H] \\
&\geq \sum_j \Pr[O_i \to O_{-i} \mid O_i \not\to H] \cdot \Pr[O_i \to j \wedge O_i \not\to H] \\
&= \Pr[O_i \to O_{-i} \mid O_i \not\to H] \cdot \sum_j \Pr[O_i \to j \wedge O_i \not\to H] \\
&\geq \Pr[O_i \to O_{-i} \mid O_i \not\to H] \cdot \mathrm{val}(O_i \mid H) \\
&\geq \Pr[O_i \to O_{-i} \mid O_i \not\to H] \cdot (1-\epsilon)\frac{\mathrm{val}(\mathrm{OPT})}{k}
\end{aligned}$$

Hence it follows that

$$\Pr[O_i \to O_{-i} \mid O_i \not\to H] \leq \frac{2\epsilon}{1-\epsilon} \leq 3\epsilon,$$

for any $\epsilon \in (0, 1/3)$, concluding the proof of the lemma.

□

B Proof of the Exclusive and Good Contribution Lemma

In this section we prove Lemma 7. Let us first recall the statement of the Lemma. We have $S = \{g_1, g_2, ..., g_{k/4}\}$ is the set of the first $k/4$ vertices chosen by the greedy algorithm. We also have that L, M, M', M'', H is a decomposition of OPT satisfying the properties of Lemma 2, Lemma 3, and Lemma 5. Recall the definitions of exclusive and good nodes for O_i, denoted E_i and G_i, from Section 3.2:

Definition 3. *We say that a vertex j is* exclusive *for $O_i \in M''$, and write $j \in E_i$, if* $\Pr[O_{-i} \to j \mid O_i \to j \wedge S \to j \wedge H \not\to j] < 48\epsilon$.

Definition 4. *We say that a vertex j is* good *for $O_i \in M''$, and write $j \in G_i$, if* $\Pr[O_i \to j \wedge S \to j \wedge H \not\to j] > \frac{1}{100}\Pr[O_i \to j \wedge H \not\to j]$.

Then Lemma 7 states that there exists a positive constant c_0 such that

$$\sum_{O_i \in M''} \sum_{j \in G_i \cap E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] > c_0 \cdot \text{val(OPT)}.$$

The proof of Lemma 7 proceeds via a sequence of claims. Let us first make some definitions for notational convenience. For sets of nodes A and B, let $\mathcal{E}^B_A$ be shorthand for the event $[A \to B]$; that is, that a node in A is connected to a node in B. We allow either A or B to be an individual node.

Claim 6. *For each $O_i \in M''$,*

$$\begin{aligned}&\sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_{O_{-i}} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] \\ &> (1 - 12\epsilon) \sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H].\end{aligned}$$

Proof. Since $O_i \in M'' \subseteq M$, we have

$$\begin{aligned}(1-\epsilon) \cdot \frac{\text{val(OPT)}}{k} &\leq \sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_{O_{-i}} \wedge \neg\mathcal{E}^j_H] \\ &\leq \sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H] \\ &\leq (1+\epsilon) \cdot \frac{\text{val(OPT)}}{k}\end{aligned}$$

from which we can conclude that $\sum_j \Pr[\mathcal{E}^j_{O_i} \wedge O_{-i} \to j \wedge \neg\mathcal{E}^j_H] < 2\epsilon\frac{\text{val(OPT)}}{k}$, and hence

$$\sum_j \Pr[\mathcal{E}^j_{O_i} \wedge O_{-i} \to j \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] < 2\epsilon\frac{\text{val(OPT)}}{k}.$$

On the other hand, we know from Lemma 4 that $v(O_i \mid S) < \frac{4}{5} \cdot \frac{\text{val(OPT)}}{k}$, and hence

$$\begin{aligned}\sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] &= \text{val}(O_i \mid H) - \text{val}(O_i \mid S, H) \\ &\geq (1-\epsilon)\frac{\text{val(OPT)}}{k} - \text{val}(O_i \mid S) \\ &\geq (\frac{1}{5} - 2\epsilon)\frac{\text{val(OPT)}}{k}\end{aligned}$$

Putting these inequalities together, and supposing $(1/5 - 2\epsilon) \geq 1/6$, we have

$$\begin{aligned}&\sum_j \Pr[\mathcal{E}^j_{O_i} \wedge O_{-i} \to j \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] \\ &< 2\epsilon\frac{\text{val(OPT)}}{k} \\ &< 12\epsilon \sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H]\end{aligned}$$

and hence

$$\begin{aligned}&\sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_{O_{-i}} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] \\ &\qquad > (1-12\epsilon)\sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H]\end{aligned}$$

as required. □

We can use the previous claim to show that much of the influence of O_i is due to nodes that are exclusive to O_i, even if we restrict our attention to nodes that are also covered by S but not covered by H.

Claim 7. *For all $O_i \in M''$,*

$$\sum_{j \in E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] > \frac{3}{4}\sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H].$$

Proof. Let λ be such that $\sum_{j \in E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] = \lambda \sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H]$. Then

$$\begin{aligned}
& (1-12\epsilon)\sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] \\
< \; & \sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_{O_{-i}} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] \\
= \; & \sum_{j \in E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_{O_{-i}} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] \\
& \quad + \sum_{j \notin E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_{O_{-i}} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] \\
< \; & \sum_{j \in E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] \\
& \quad +(1-48\epsilon)\sum_{j \notin E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] \\
= \; & \lambda \sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] \\
& \quad +(1-48\epsilon)(1-\lambda)\sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H].
\end{aligned}$$

It follows that $(1-12\epsilon) < \lambda + (1-48\epsilon)(1-\lambda)$, from which we conclude $\lambda \geq 3/4$ as required. □

We next claim that much of the influence of each O_i is captured by nodes that are good for O_i. We will actually show something stronger: that this is true even if we restrict our attention only to good nodes that are also exclusive to O_i.

Claim 8. *There exists a constant $c_1 > 0$ such that, for each i,*

$$\sum_{j \in G_i \cap E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H] \geq c_1 \sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H].$$

Proof. We first claim that, for each i,

$$\frac{3}{20}\sum_{j \in E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H] \leq \sum_{j \in E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H].$$

To see this, note

$$\begin{aligned}
\sum_{j \in E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H] &\leq \sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H] \\
&\leq 5\sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] \\
&\leq 5 \cdot \frac{4}{3}\sum_{j \in E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H]
\end{aligned}$$

as required, where the second inequality follows because $\text{val}(O_i \mid S, H) \leq \frac{4}{5}(1-\epsilon)\frac{\text{val(OPT)}}{k} \leq \frac{4}{5}\text{val}(O_i \mid H)$.

Now suppose $\sum_{j \in E_i \cap G_i} \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H] = \lambda \sum_{j \in E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H]$. We then have

$$
\begin{aligned}
&\frac{3}{20} \sum_{j \in E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H] \\
&\leq \sum_{j \in E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] \\
&\leq \sum_{j \in E_i \cap G_i} \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H] + \sum_{j \in E_i \setminus G_i} \frac{1}{100} \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H] \\
&= \lambda \sum_{j \in E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H] + \frac{1}{100}(1-\lambda) \sum_{j \in E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H]
\end{aligned}
$$

from which we conclude $\frac{3}{20} \leq \lambda + \frac{1}{100}(1-\lambda)$, which implies $\lambda \geq \frac{14}{99}$.

But now, since $\sum_{j \in E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] \geq \frac{3}{4} \sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H]$ from the previous claim, we get

$$
\begin{aligned}
\sum_{j \in G_i \cap E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H] &\geq \frac{14}{99} \sum_{j \in E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H] \\
&\geq \frac{14}{99} \sum_{j \in E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] \\
&\geq \frac{14}{99} \cdot \frac{3}{4} \sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] \\
&\geq \frac{14}{99} \cdot \frac{3}{4} \cdot \frac{1}{5} \sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H],
\end{aligned}
$$

where the last inequality follows because $\text{val}(O_i \mid S, H) \leq \frac{4}{5}(1-\epsilon)\frac{\text{val}(\text{OPT})}{k} \leq \frac{4}{5}\text{val}(O_i \mid H)$. This yields the desired result, with constant factor $c_1 = \frac{14}{99} \cdot \frac{3}{4} \cdot \frac{1}{5} > \frac{1}{50}$. □

We can now complete the proof of Lemma 7. We have

$$
\begin{aligned}
&\sum_i \sum_{j \in G_i \cap E_i} \Pr[\mathcal{E}^j_{O_i} \wedge \mathcal{E}^j_S \wedge \neg\mathcal{E}^j_H] \\
&\geq \sum_i \sum_{j \in G_i \cap E_i} \frac{1}{100} \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H] \\
&\geq \frac{c_1}{100} \sum_i \sum_j \Pr[\mathcal{E}^j_{O_i} \wedge \neg\mathcal{E}^j_H] \\
&> \frac{c_1}{100} \sum_i (1-\epsilon)\frac{\text{val}(\text{OPT})}{k} \\
&\geq \frac{c_1}{100} \cdot \frac{k}{3}(1-\epsilon)\frac{\text{val}(\text{OPT})}{k} \\
&= c_0 \text{val}(\text{OPT})
\end{aligned}
$$

for constant $c_0 = \frac{c_1}{100} \cdot \frac{1}{3} \cdot (1-\epsilon) > \frac{c_1}{400}$, completing the proof of Lemma 7.