Open access • Journal Article • DOI:10.1109/TNET.2013.2290714

Cost-effective viral marketing for time-critical campaigns in large-scale social networks — Source link

Thang N. Dinh, Huiyuan Zhang, Dzung T. Nguyen, My T. Thai

Institutions: University of Florida

Published on: 01 Dec 2014 - IEEE ACM Transactions on Networking (IEEE)

Topics: Viral marketing and Network theory

Cost-effective Viral Marketing for Time-critical Campaigns in Large-scale Social Networks

Thang N. Dinh, Huiyuan Zhang, Dung T. Nguyen, and My T. Thai

***Abstract*—Online social networks (OSNs) have become one of the most effective channels for marketing and advertising. Since users are often influenced by their friends, "word-of-mouth" exchanges, so-called viral marketing, in social networks can be used to increase product adoption or widely spread content over the network. The common perception of viral marketing about being cheap, easy, and massively effective makes it an ideal replacement of traditional advertising. However, recent studies have revealed that the propagation often fades quickly within only few hops from the sources, counteracting the assumption on the self-perpetuating of influence considered in literature. With only limited influence propagation, is massively reaching customers via viral marketing still affordable? How to economically spend more resources to increase the spreading speed?**

We investigate the cost-effective massive viral marketing problem, taking into the consideration the limited influence propagation. Both analytical analysis based on power-law network theory and numerical analysis demonstrate that the viral marketing might involve costly seeding. To minimize the seeding cost, we provide mathematical programming to find optimal seeding for medium-size networks and propose VirAds, an efficient algorithm, to tackle the problem on large-scale networks. VirAds guarantees a relative error bound of $O(1)$ from the optimal solutions in power-law networks and outperforms the greedy heuristics which realizes on the degree centrality. Moreover, we also show that, in general, approximating the optimal seeding within a ratio better than $O(\log n)$ is unlikely possible.

***Index Terms*—Approximation algorithm, hardness proof, social media, influence propagation, power-law networks;**

I. Introduction

Digitizing real world connections, online social networks (OSNs) such as Twitter and Facebook have been steadily growing. Two-third of everyone online is using social networks with more than one billion active Facebook users [1], 200 million twitters, 40 millions Google+ subscribers and so on. Social network sites such as Facebook and Youtube are often among top-ten visited websites on the Internet [2]. Much like real-world social networks, OSNs inherent the viral property in which information can spread and disseminate widely into networks via 'word-of-mouth' exchanges. They are effective channels to increase brand awareness, encourage discussion on improving products, and recruit new employees. Notable examples include the recent unrest in many Arab countries which are triggered by Facebook shared posts [3]; the customer outreach of Toyota on Twitter to repair its image after the massive safety recalls of its vehicles [4], and many others. Despite the huge economic and political impact, viral marketing in large scale OSNs is not well understood due to the extremely large numbers of users and complex structures of social links.

A major portion of viral marketing research has been devoted to the question of efficiently targeting a set of *influential nodes* in order to spread information widely into the network [5]–[7]. Two essential components to address the question are the diffusion models and the algorithms to select an initial set of nodes, called seeding. For a social network represented as a graph, a diffusion model defines the stochastic process that specifies how influence is propagated from the seeding to their neighbors, and further. In [5], Kempe et al. proposed two basic diffusion models, namely *independent cascade* and *linear threshold* models. These two models and their extensions set the foundation to almost all existing algorithms to find seeding in social networks [6]–[9].

However, all the mentioned models and algorithms ignore one important aspect of influence propagation in the real world, i.e., influence propagation often happens only within a close proximity of the seeding. For example, a study of Flickr [10] reveals that the typical chain length is less than four; another study of Leskovec et. al. [11] suggests that social influence happens on the level of direct friends. Moreover, shared information in social networks such as Facebook can usually be seen only by friends or friends of friends i.e. the propagation is basically limited within two hops from the source. Thus it is often sufficient to consider the propagation within a few hops of the seeding. When the influence only propagates locally, is massively reaching customers via viral marketing still affordable? Also, can we speed up the information spreading for time-critical applications such as political campaigns?

We formulate a new optimization problem, called *the cost-effective, fast, and massive viral marketing* (CFM) problem. The problem seeks for a minimal cost seeding, measured as the number of nodes, to massively and quickly spread the influence to the whole network (or a large segment of the network). The new aspect in our model is that the influence is limited to the nodes that are within d hops from the seeding for some constant $d \geq 1$. In

This work is partially supported by the DTRA YIP grant number HDTRA1-09-1-0061 and the NSF CAREER Award number 0953284.

T. N. Dinh, Huiyuan Zhang, D. T. Nguyen, and M. T. Thai are with the Department of Computer & Information Science & Engineering, University of Florida, Gainesville, FL, 32611.
E-mail: {tdinh, huiyuan, dtnguyen, mythai}@cise.ufl.edu.

T. N. Dinh is with Department of Computer Science, Virginia Commonwealth University. Email: tndinh@vcu.edu

other words, the influence is forced to spread to the whole networks within d propagation rounds. Hence, adjusting d gives us an important ability to control how fast the spread of influence within a network. Unfortunately, the huge magnitude of OSN users and data available on OSNs poses a substantial challenge to control how information can quickly spread out to the whole network.

In this paper, we develop solutions to the CFM problem and address the above two questions. Our contributions are summarized as follows:

- Our first finding shows that the seeding for fast and massive spreading must contain a non-trivial fraction of nodes in the networks, which is cost-prohibitive for large social networks. This is confirmed by both our theoretical analysis based on the power-law model in [12] and our extensive experiments.
- We propose VirAds, a scalable algorithm to find a set of minimal seeding to expeditiously propagate the influence to the whole network. VirAds outperforms the greedy heuristics based on well-known degree centrality and scales up to networks of hundred of million links. We prove that the algorithm guarantees a relative error bound of $O(1)$, assuming that the network is power-law.
- We show how hard to obtain a near optimal solution for CFM by proving the impossibility to approximate the optimal solution within a ratio better than $O(\log n)$.

Related Work. Viral marketing can be thought of as a diffusion of information about the product and its adoption over the network. Kempe et al. [5], [6] formulated the influence maximization problem as an optimization problem. They showed the problem to be NP-complete and devised an $(1-1/e-\epsilon)$ approximation algorithm. A major drawback of their algorithm is that the accuracy ϵ, and efficiency depends on the number of sampling times to simulate the propagation process. Later, Leskovec et al. [13] study the influence propagation in a different perspective in which they aim to find a set of nodes in networks to detect the spread of virus as soon as possible. They improve the simple greedy method with the lazy-forward heuristic (CELF), which is originally proposed to optimize submodular functions in [14]. The greedy algorithm is furthered improved by Chen et al. [8] by using an influence estimation. For the Linear Threshold model, Chen et al. [15] proposed to use local directed acyclic graphs (LDAG) to approximate the influence regions of nodes.

More recently, Goyal et al. [16] studied the problem of MINTIME in which given a coverage threshold η and a budget threshold k, the task is to find a seed set of size at most k such that by activating it, at least η nodes are activated in expectation in minimum possible time. In MINTIME, the objective is to minimize the influence timing, when in our case, time is given as a deadline constraint. Chen et al. in [17] extends the IC and LT models to study the time-critical influence maximization problem. The goal is to identify k nodes such that activating those nodes maximizes the expectation of the number of activated nodes within a given deadline. This objective is different from ours that aims to minimize the seeding cost. In addition, both [16] and [17] study the influence propagation based on the IC and LT models, while our study is based on a deterministic threshold model.

Influence propagation with limited number of hops is first considered in Wang et al. [18] and Feng et al. [19] for the special case $d = 1$ and $\rho = 1/2$. We note that none of the mentioned approaches handled large-scale social networks of millions of nodes as we shall study in Section VI.

Organization. We introduce the limited hop influence model and the cost-effective, massive and fast propagation problem (CFM) in Section II. In Section III, we answer the question on the seeding cost by analyzing the propagation process on power-law networks. We present VirAds, a scalable algorithm to find a minimal seeding for the CFM problem in Section IV. The hardness of finding a cost-effective seeding is addressed in Section V. Finally, we perform extensive experiments on large social networks such as Facebook and Orkut to confirm the efficiency of our proposed algorithm and analyze the results to give new observations to information diffusion process in networks.

II. Problem Definitions

We are given a *social network* modeled as an undirected graph $G = (V, E)$ where the vertices in V represent users in the network and the edges in E represent social links between users. We use n and m to denote the number of vertices and edges, respectively. The set of neighbors of a vertex $v \in V$ is denoted by $N(v)$ and we denote by $d(v) = |N(v)|$ the degree of node v.

Next we specify the diffusion model that governs the process of influence propagation. Existing diffusion models can be categorized into two main groups [5]:

- *Threshold model.* Each node v in the network has a threshold $t_v \in [0, 1]$, typically drawn from some probability distribution. Each connection (u, v) between nodes u and v is assigned a weight $w(u, v)$. For a node v, let $F(v)$ be the set of neighbors of v that are already influenced. Then v is influenced if $t_v \leq \sum_{u \in F(v)} w(u, v)$.
- *Cascade model.* Whenever a node u is influenced, it is given a single chance to activate each of its neighbor v with a given probability $p(u, v)$.

Most viral marketing papers assume that the probabilities $p(u, v)$ or weights $w(u, v)$ and thresholds t_v are given as a part of the input. However, they are generally not available and inferring those probabilities and thresholds has remained a non trivial problem [20]. Thus in addition to the bounded propagation hop, we use a simplified variation of the linear threshold model in which a vertex is activated if a fraction ρ of its neighbors are active as follows.

Locally Bounded Diffusion Model. Let $R_0 \subset V$ be the subset of vertices selected to initiate the influence propagation, which we call the *seeding*. We also call a vertex $v \in R_0$ a seed. The propagation process happens in round, with all vertices in R_0 are influenced (thus active in adopting the behavior) at round $t = 0$. At a particular round $t \geq 0$, each vertex is either active (adopted the behavior) or inactive and each vertex's tendency to become active increases when more of its neighbors become active. If an inactive vertex u has more than $\lceil \rho\ d(u) \rceil$ active neighbors at round t, then it becomes active at round $t+1$, where ρ is the *influence factor* as discussed later. The process goes on for a maximum number of d rounds and a vertex once becomes active will remain active until the end. We say an initial set R_0 of vertices to be a *d-seeding* if R_0 can activate at least a fraction $0 < \tau \leq 1$ of vertices in the networks within at most d rounds.

The influence factor $0 < \rho < 1$ is a constant that decides how widely and quickly the influence propagates through the network. Influence factor ρ reflects real-world factors such as how easy to share the content with others, or some intrinsic benefit for those who initially adopt the behavior. In case $\rho = 1/2$ the model is also known as the *majority* model that has many application in distributed computing, voting system [21], etc.

In addition, we denote the desired fraction of nodes to be activated after d rounds by τ, the *coverage constant*. Given the diffusion model and constants $0 < \rho < 1$, $0 < \tau \leq 1$, and $d \in \mathbb{Z}^+$, we define the *Cost-effective, Fast, and Massive viral marketing (CFM)* problem as follows.

Definition 1 (CFM Problem): Given an undirected graph $G = (V, E)$ modeling a social network, find in V a minimum subset of nodes that can activate at least $\tau|V|$ nodes within at most d rounds.
In other words, the problem asks for a minimum size d-seeding.

Generalization. The diffusion model can be generalized in several ways. For example, the model can be extended naturally to cover directed and weighted networks or specify different influence factor ρ_v for each node $v \in V$. Furthermore, each node v can be associated with a positive number θ_v that represents the cost to add v into the seeding. For simplicity we stick with the current model to avoid setting parameters during the experiments. Nevertheless, major results such as the approximation ratio of the VirAds algorithm in Section IV or the hardness of approximation result in Section V still hold for the generalized models.

III. Cost of Massive Marketing

In this section, we give a negative answer for the first question in the introduction about the initial seeding cost. We exploit the power-law topology found in most social networks [22]–[24] to demonstrate that when the propagation hop is limited, a large number of seeding nodes is needed to spread the influence throughout the network. The size of seeding is proved to be a constant fraction of the number of vertices n, which is prohibitive for large social networks of millions of nodes. We first summarize the well-known power-law model in [25]; then we use the model to prove the prohibitive seeding cost for the CFM problem.

A. Power-law Network Model.

Many complex systems of interest including OSNs are found to have the degree distributions approximately follows the power laws [22]–[24]. That is the fraction of nodes in the network having k connections to other nodes is proportional to $k^{-\gamma}$, where γ is a parameter whose value is typically in the range $2 < \gamma < 3$. Those networks have been used in studying different aspects of the scale-free networks [25]–[27]. We follow the $P(\alpha, \gamma)$ power-law model in [25] in which the number of vertices of degree k is $\lfloor \frac{e^\alpha}{k^\gamma} \rfloor$ where e^α is the normalization factor. We note that our result do not rely on the assumption in the $P(\alpha, \gamma)$ model that vertices are connected at random. Thus our result in this section holds for all random graphs in the model.

We can deduce that the maximum degree in a $P(\alpha, \gamma)$ network is $e^{\frac{\alpha}{\gamma}}$ (since for $k > e^{\frac{\alpha}{\gamma}}$, the number of edges will be less than 1). The number of vertices and edges are

$$n = \sum_{k=1}^{e^{\frac{\alpha}{\gamma}}} \frac{e^\alpha}{k^\gamma} \approx \begin{cases} \zeta(\gamma)e^\alpha & \text{if } \gamma > 1 \\ \alpha e^\alpha & \text{if } \gamma = 1 \\ \frac{e^{\frac{\alpha}{\gamma}}}{1-\gamma} & \text{if } \gamma < 1 \end{cases},$$

$$m = \frac{1}{2} \sum_{k=1}^{e^{\frac{\alpha}{\gamma}}} k \frac{e^\alpha}{k^\gamma} \approx \begin{cases} \frac{1}{2}\zeta(\gamma-1)e^\alpha & \text{if } \gamma > 2 \\ \frac{1}{4}\alpha e^\alpha & \text{if } \gamma = 2 \\ \frac{1}{2}\frac{e^{\frac{2\alpha}{\gamma}}}{2-\gamma} & \text{if } \gamma < 2 \end{cases} \quad \text{(III.1)}$$

where $\zeta(\gamma) = \sum_{i=1}^{\infty} \frac{1}{i^\gamma}$ is the Riemann Zeta function [25] which converges for $\gamma > 1$ and diverges for all $\gamma \leq 1$. Without affecting the conclusion, we will use real numbers instead of rounding down to integers. The error terms are sufficiently small and can be bounded in our proofs.

While the scale of the network depends on α, the parameter γ decides the connection pattern and many other important characterizations of the network. For instance, the larger γ, the sparser and the more "power-law" the network is. Hence, the parameter γ is often regarded as the characteristic constant for scale-free networks.

B. Prohibitive Seeding Costs

We prove that the seeding must contain at least $\Omega(n)$ vertices if the propagation is locally bounded. The result is stated in the following theorem.

Theorem 1: Given a power-law network $G \in P(\alpha, \gamma)$, with $\gamma > 2$ and constants $0 < \rho < 1$ and $0 < \tau \leq 1$, any d-seeding is of size at least $\Omega(n)$.

Proof: The proof consists of two parts. In the first part, we show that the volume i.e. the total degree of vertices, of any d-seeding must be $\Omega(m)$. In the second part, we prove that any subset of vertices $S \subset V$ with volume $\text{vol}(S) = \Omega(m)$ in a power-law network with power-law exponent $\gamma > 2$, will imply that $|S| = \Omega(n)$. Thus, the theorem follows.

In the first part, we consider two separate cases

Case $\rho > \frac{1}{2}$: Let $S = R_0$ be the optimal solution for the CFM problem on $G = (V, E)$, and $S = R_0, R_1, R_2, \ldots, R_d$ are vertices that become active at round $0, 1, 2, \ldots, d$, respectively. For each $1 \leq t \leq d$ the following inequality holds.

$$|\phi(R_t, \bigcup_{i=0}^{t-1} R_i)| \geq \rho \text{vol}(R_t) \quad \text{(III.2)}$$

where $\phi(A, B)$ denotes the set of edges connecting one vertex in A to one vertex in B and $\text{vol}(R_t)$ denotes the total degree of vertices in R_t. The inequality means that at least a fraction ρ among the edges incident to the vertices activated in round t must be incident with active vertices in the previous rounds.

Sum up all inequalities in (III.2) for $t = 1..d$, we have

$$\sum_{t=1}^{d} |\phi(R_t, \bigcup_{i=0}^{t-1} R_i)| \geq \rho \text{vol}(\bigcup_{i=1}^{d} R_i) \tag{III.3}$$

Note that the left side can be written as

$$\frac{1}{2}\left(\sum_{t=1}^{d} |\phi(R_t, \bigcup_{i=0}^{t-1} R_i)| + \sum_{i=0}^{d-1} |\phi(\bigcup_{t=i+1}^{d} R_t, R_i)|\right)$$
$$\leq \frac{1}{2}|\phi(\bigcup_{t=0}^{d} R_t, \bigcup_{t=0}^{d} R_t)| \leq \frac{1}{2}\text{vol}(\bigcup_{t=0}^{d} R_t) \tag{III.4}$$

From Eqs. III.3 and III.4, we have

$$\frac{1}{2}\text{vol}(\bigcup_{t=0}^{d} R_t) \geq \rho \text{vol}(\bigcup_{t=1}^{d} R_t)$$
$$\Rightarrow \text{vol}(R_0) \geq \frac{2\rho - 1}{\rho} \text{vol}(\bigcup_{t=0}^{d} R_t) \tag{III.5}$$

Since there must be at least τn activated vertices after d rounds, we have $|\bigcup_{t=0}^{d} R_t| \geq \tau n$. Then, by III.1

$$\text{vol}(\bigcup_{t=0}^{d} R_t) \geq |\bigcup_{t=0}^{d} R_t| \geq \quad \tau n = \frac{\zeta(\gamma - 1)}{2\zeta(\gamma)} m = \Omega(m)$$

Hence, when $\rho > 1/2$, $\text{vol}(R_0) = \Omega(m)$ for any d-seeding R_0.

Case $\rho \leq \frac{1}{2}$: We say that an *edge is active if it is incident to at least one active vertex*. At round $t = 0$, there are at most $\text{vol}(R_0)$ *active edges*, those who are incident to R_0. Eq. III.2 implies that the number of active edges in each round increases at most ρ^{-1} times. After d rounds, the number of active edges will be bounded by $\text{vol}(R_0) \times \rho^{-d}$. Since the number of active edge after d rounds is at least $\Omega(m)$, we have

$$\text{vol}(R_0) \geq \rho^{-d} \text{vol}(\bigcup_{t=0}^{d} R_t) = \Omega(m).$$

This completes the first part of the proof.

In the second part of the proof, we show that if a subset $S \subset V$ has $\text{vol}(S) = \Omega(m)$, then $|S| = \Omega(n)$ whenever the power-law exponent $\gamma > 2$. Assume that $\text{vol}(S) \geq cm$, for some positive constant c. The size of S is minimum when S contains only the highest degree vertices of V. Let k_0 be the minimum degree of vertices in S in that extreme case, by Eq. III.1 we have

$$cm = \frac{c}{2} \sum_{k=1}^{e^{\frac{\alpha}{\gamma}}} k \frac{e^{\alpha}}{k^{\gamma}} \leq \text{vol}(S) \leq \frac{1}{2} \sum_{k=k_0}^{e^{\frac{\alpha}{\gamma}}} k \frac{e^{\alpha}}{k^{\gamma}}$$

Simplify two sides, we have

$$\sum_{k=1}^{k_0 - 1} \frac{1}{k^{\gamma - 1}} \leq (1 - c) \sum_{k=1}^{e^{\frac{\alpha}{\gamma}}} \frac{1}{k^{\gamma - 1}} = (1 - c)\zeta(\gamma - 1)$$

Since, the zeta function $\zeta(\gamma - 1)$ converges for $\gamma > 2$, there exists a constant $k_{\rho,\gamma}$ that depends only on ρ and γ that satisfies

$$\sum_{k=1}^{k_{\rho,\gamma}} \frac{1}{k^{\gamma - 1}} > (1 - c)\zeta(\gamma - 1)$$

Obviously, we have $k_0 \leq k_{\rho,\gamma}$. Thus, the number of vertices that are in S is at least

$$\sum_{k=k_{\rho,\gamma}}^{e^{\frac{\alpha}{\gamma}}} \frac{e^{\alpha}}{k^{\gamma}} = (1 - \sum_{k=1}^{k_{\rho,\gamma}} \frac{1}{k^{\gamma}})n = \Omega(n)$$

We have the last step because the sum $\sum_{k=1}^{k_{\rho,\gamma}} \frac{1}{k^{\gamma}}$ is bounded by a constant since $k_{\rho,\gamma}$ is a constant. ■

In both cases $\rho > 1/2$ and $\rho \leq 1/2$, the size of a d-seeding set is at least $\Omega(n)$. However, we can see a clear difference in the propagation speed with respect to d between two cases. When $\rho < 1/2$, the number of active edges can increase exponentially (but is still bounded if d is a constant) and, it is likely that the number of active vertices also exponentially increases. In contrast, when $\rho > 1/2$, exploding in the number of active edges (and hence active vertices) is impossible as the total number of active edges after d rounds is bounded by $\frac{\rho}{2\rho - 1}$ volume of the initial d-seeding, regardless of the value of d.

IV. Cost-Effective & Expeditious Social Marketing Algorithm

In order to understand the influence propagation when the number of propagation hops is bounded, we propose VirAds, an efficient algorithm for the CFM problem. With the huge magnitude of OSN users and data available on OSNs, scalability becomes the major problem in designing algorithm for CFM. VirAds is scalable to network of hundred of millions links and provides high quality solutions in our experiments.

Before presenting VirAds, we consider a natural greedy algorithm for the CFM problem in which the vertex that can activate the largest number of inactive vertices within d hops is selected in each step. This greedy algorithm is unlikely to perform well in practice for following two reasons. First, at early steps, when not many vertices are selected, every vertex is likely to activate only itself after being chosen as a seed. Thus, the algorithm cannot distinguish between good and bad seeds. Second, the algorithm suffers serious scalability problems at later selection steps. To select a vertex, the algorithm has to evaluate for each vertex v how many vertices will be activated after adding v to the seeding, e.g. by invoking an $O(m+n)$ Breadth-First Search procedure rooted at v. In the worst-case when $O(n)$ vertices are needed to evaluate, this alone can take $O(n(m + n))$. Moreover, as shown in the previous section, the seeding size can be easily $\Omega(n)$; thus, the worst-case running time of the naive greedy algorithm is $O(n^2(m+n))$, which is prohibitive for large-scale networks.

As shown in Algorithm 1, our VirAds algorithm overcomes the mentioned problems in the naive greedy by favoring the vertex which can activate the most number of *edges* (indeed, it also considers the number of active neighbor around each vertex). This avoids the first problem of the naive greedy algorithm. At early steps, the algorithm behaves similar to the degree-based heuristics that favors vertices with high degree. However, when a certain number of vertices are selected, VirAds will make the selection based on the information within d-hop neighbor around the considered vertices rather than only one-hop neighbor as in the degree-based heuristic.

Algorithm 1: VirAds - Viral Advertising in OSNs

```
Input: Graph $G = (V, E)$, $0 < \rho < 1$, $d \in \mathbb{N}^+$
Output: A small $d$-seeding
$n_v^{(e)} \leftarrow d(v), n_v^{(a)} \leftarrow \rho \cdot d(v), r_v \leftarrow d+1, v \in V$;
$r_v^{(i)} = 0, i = 0..d, P \leftarrow \emptyset$;
while number of active nodes is less than $\tau|V|$ do
    repeat
        $u \leftarrow$ argmax$_{v \notin P}$ $\{n_v^{(e)} + n_v^{(a)}\}$;
        Recompute $n_u^{(e)}$ as the number of
        new active edges after adding $u$.
    until $u = argmax_{v \notin P}$ $\{n_v^{(e)} + n_v^{(a)}\}$;
    $P \leftarrow P \cup \{u\}$;
    Initialize a queue: $Q \leftarrow \{(u, r_u)\}$;
    $r_u \leftarrow 0$;
    foreach $x \in N(u)$ do
        $n_x^{(a)} \leftarrow \max\{n_x^{(a)} - 1, 0\}$;
    while $Q \neq \emptyset$ do
        $(t, \tilde{r}_t) \leftarrow Q.pop()$;
        foreach $w \in N(t)$ do
            foreach $i = r_t$ to $\min\{\tilde{r}_t - 1, r_w - 2\}$ do
                $r_w^{(i)} = r_w^{(i)} + 1$;
                if $r_w^{(i)} \geq \rho \cdot d_w$ then
                    if $(r_w \geq d) \wedge (i + 1 < d)$ then
                        foreach $x \in N(w)$ do
                            $n_x^{(a)} \leftarrow \max\{n_x^{(a)} - 1, 0\}$;
                    $r_w = i + 1$;
                    if $(r_w < d) \wedge (w \notin Q)$ then
                        $Q.push(w, r_w)$;
Output $P$;
```

The scalability problem is tackled in VirAds by efficiently keeping track of the following measures for each vertex v.

- $\mathbf{r_v}$: the round in which v is activated
- $\mathbf{n_v^{(e)}}$: The number of new active edges after adding v into the seeding
- $\mathbf{n_v^{(a)}}$: The number of extra active neighbors v needs in order to activate v
- $\mathbf{r_v^{(i)}}$: The number of activated neighbors of v up to round i where $i = 1..d$.

Given those measures, VirAds selects in each step the vertex u with the highest *effectiveness* which is defined as $n_u^{(e)} + n_u^{(a)}$. After that, the algorithm needs to update the measures for all the remaining vertices.

We use an effective updating strategy, referred as "*Smart-update*", to maintain node activation status incrementally. In our experiments in Appendix B, the "Smart-update" strategy reduce the running time of VirAds more than 200 times. Except for $n_v^{(e)}$, all other measures can be effectively kept track of in only $O((m+n)d)$ during the whole algorithm. When a vertex u is selected, it causes a chain-reaction and activate a sequence of vertices or lower the rounds in which vertices are activated. New activated vertices together with their active rounds are successively pushed into the queue Q for further updating much like what happens in the Bellman-Ford shortest-paths algorithm. Every time we pop a vertex v from Q. If r_v, the current active round of v, is different from $\tilde{r}_v$, the previous active round of v when v is pushed into Q, we update for each neighbor w of v the values of r_w and $r_w^{(i)}$. If any neighbor w of v changes its active round and w is not in Q, we push w into Q for further update. The update process stops when Q is empty. Note that for each node $u \in V$, changing of r_u can cause at most d update for $r_w^{(.)}$ where w is a neighbor of u. For all neighbors of u, the total number of update is, hence, $O(d \cdot d(u))$. Thus, the total time for updating $r_w^{(.)}$ $\forall w \in V$ in VirAds will be at most $O((m+n) \cdot d)$.

To maintain $n_v^{(e)}$, the easiest approach is to recompute all $n_v^{(e)}$. This approach, called *Exhaustive Update*, is extremely time-consuming as discussed in the naive greedy. Instead, we only update $n_v^{(e)}$ when "necessary". In details, vertices are stored in a max priority queue in which the priority is their *effectiveness*. In each step, the vertex u with the highest effectiveness is extracted and $n_u^{(e)}$ is recomputed. If after updating, u still has the highest effectiveness, u is then selected. Otherwise, u is pushed back to the priority queue, and the new vertex with the highest effectiveness is considered, and so on. This strategy is similar to the accelerated greedy proposed in [14] for maximizing submodular functions and the lazy-forward heuristic (CELF) in [13]. Note that since the CFM problem does not exhibit submodular property, this strategy does not guarantee the selection of the node with highest effectiveness. Nevertheless, the experiments suggest that this strategy outperforms the Exhaustive Update in terms of running time and retains high-quality solutions at the same time.

While we cannot give a better bound than $O(n^2(m+n))$ for the worst case, the VirAds algorithm is expected to run in an $O(s \cdot d(m+n))$ time, where s is the number of seed nodes in the solution and d is the number of propagation round. In our experiments, VirAds reduces the running time up to several thousand times, while retaining the high quality solutions.

Approximation Ratio for Power-law Networks. The CFM problem can be easily shown to be NP-hard by a reduction from the set cover problem. Thus, we are left with two choices: designing heuristics which have no worst-case performance guarantees or designing approximation algorithms which can guarantee the produced solutions are within a certain factor from the optimal. Formally, a β-approximation algorithm for a minimization (maximization) problem always returns solutions that are at most β times larger (smaller) than an optimal solution.

Unfortunately, there is unlikely an approximation algorithm with factor less than $O(\log n)$ as shown in next section. However, under the assumption that the degree distribution follows a power-law, our VirAds, or any algorithm that gives feasible solutions for CFM, is an approximation algorithm for CFM with a constant factor. This follows directly from the result in previous section that the optimal solution has size at least $\Omega(n)$ in power-law networks. Thus, the ratio between the VirAds's solution and the optimal solution is bounded by a constant.

V. Hardness of Identifying Seeding with Guarantees

This section provides the hardness of approximating the optimal solutions of the CFM problem, i.e., the impossibility of finding near-optimal solutions in polynomial time. In the previous section, we can obtain $O(1)$ approximation algorithms for CFM when the network is power-law. In contrast, we can modify slightly the main reduction in [8] to show that the CFM problem cannot be approximated within a factor $\Omega(2^{\log^{1-\epsilon} n})$ for $d \geq 6$ and $\epsilon > 0$. This implies that unlike Linear Threshold and Independent Cascade models [5], the locally bounded diffusion model in this paper is non-submodular in general.

For the cases $d \leq 6$, we show below that there is no algorithm that can approximate the problem within a factor less than $\Omega(\log n)$. We first prove the hardness for the case when $d = 1$ and $\tau = 1$. Then we present the hardness for the cases $1 < d \leq 6$.

A. One-hop CFM

We prove that the CFM problem cannot be approximated within a factor $\ln \Delta - (\ln \ln \Delta)$ in graphs of maximum degree Δ, unless P=NP. The proof uses a gap-reduction from an instance of the *Bounded Set Cover* problem (SC$_B$) to an instance of CFM problem whose degrees are bounded by $B' = B \text{ poly} \log B$. The proof can be found in the conference version of the paper [28].

Theorem 2: [28] When $d = 1$, it is NP-hard to approximate the CFM problem in graphs with degrees bounded by B' within a factor of $\ln B' - c_1 \ln \ln B'$, for some constant $c_1 > 0$.

Similarly, with appropriate setting in Feige's construction [29], we obtain the following hardness result regarding the network size n. The proof is given in [30] or Appendix A.

Theorem 3: [30] For any $\epsilon > 0$, the CFM problem, when $d = 1$, cannot be approximated within a factor $(\frac{1}{2} - \epsilon) \ln n$, unless NP $\subset$ DTIME$(n^{O(\log \log n)})$.

Note that Theorems 2 and 3 are incomparable in general. Let Δ be the maximum degree, Theorem 2 implies the hardness of approximation with factor $(1 - \epsilon) \ln \Delta$, which is larger than $(\frac{1}{2} - \epsilon) \ln n$ if $\Delta \approx n$, but smaller than $(\frac{1}{2} - \epsilon) \ln n$ when $\Delta < \sqrt{n}$. In addition, the Theorem 3 uses a stronger assumption than that in Theorem 2.

B. Multiple-hop CFM

We now present a gap reduction from the (d-hop) CFM problem to the one-hop CFM problem with $d \geq 2$. The hardness result follows immediately by the Theorem 2 in the previous section.

Given a graph $G = (V, E)$ as an instance of the CFM problem. We will construct an instance $G' = (V', E')$ of the CFM problem as follows (and as illustrated in Fig. 2). We add $c(\rho)$ vertices $w_1, w_2, \ldots, w_{c(\rho)}$, called flashpoints, where $c(\rho) = \min\{t \in \mathbb{N} \mid \frac{t-1}{t+1} \leq \rho < \frac{t}{t+1}\}$. These vertices will be selected at the beginning to kick off the activation of other nodes. Furthermore, each "flashpoint" w_p is connected to a dummy vertex z_p.

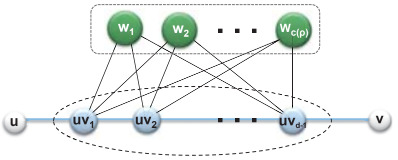


Fig. 1. The transmitter gadget.

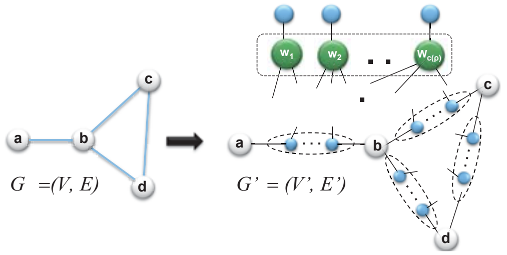


Fig. 2. Gap-reduction from one-hop CFM to d-hop CFM.

Replace each edge $(u, v) \in E$ by a gadget called transmitter. The transmitter connecting vertex u and v is a chain of $d-1$ path, named uv_1 to uv_{d-1}. The vertex u is connected to uv_1, uv_1 is connected to uv_2 and so on, vertex uv_{d-1} is connected to v. Each vertex uv_i, $i = 1..d-1$ is connected to all flashpoints. An example for transmitter is shown in Fig. 1. The transmitter is designed so that if all flashpoints and vertex u are selected at the beginning, then vertex uv_{d-1} will be activated after $d-1$ rounds. Hence, the number of activated neighbors of v after $d-1$ rounds will equal the number of selected neighbors of v in the original graph.

Finally, we replace each edge (w_p, z_p) by a transmitter. In order to activate all dummy vertices z_p after d rounds, assume, w.l.o.g., that all flashpoints must be selected in an optimal solution. The following lemma follows directly from the construction.

Lemma 1: Every solution of size k for the one-hop ($d = 1$) CFM problem in G induces a solution of size $k + c(\rho)$ for the d-hop CFM problem in G'.

On another direction, we also have the following lemma.

Lemma 2: An optimal solution of size k' for the d-hop CFM problem induces a size $k' - c(\rho)$ solution for the one-hop CFM problem in G.

Proof: For a transmitter connecting u to v, if the solution of the d-hop CFM problem contains any of the intermediate vertices $uv_1, \ldots, uv_{d-1}$, we can replace that vertex in the solution with either u or v to obtain a new solution of same size (or less). Hence, we can assume, w.l.o.g., that none of the intermediate vertices are selected. Therefore, all flashpoints must be selected in order to activate the dummy vertices. It is easy to see that the solution of d-hop CFM excluding the flashpoints will be a solution of one-hop CFM in G with size $k' - c(\rho)$. ■

Note that the number of vertices in G' is upper-bounded by dn^2 i.e. $\ln |V'| < 2ln|V| + lnd$. Thus, using the same arguments used in the proof of Theorem 3, we can show that a $(\frac{1}{4} - \epsilon)\ln n$ approximation algorithm algorithm lead to a $(\frac{1}{2} - \epsilon)\ln n$ approximation algorithm for the one-hop CFM problem (contradicts Theorem 3).

Theorem 4: The CFM problem cannot be approximated within $(\frac{1}{4} - \epsilon)\log n$ for $d \geq 1$, unless NP $\subset$ DTIME($n^{O(\log\log n)}$)

VI. Empirical Study

In this section we perform experiments on OSNs to show the efficiency of our algorithms in comparison with simple degree centrality heuristic and study the trade-off between the number of times the information is allowed to propagate in the network and the seeding size.

A. Comparing to Optimal Seeding

One advantage of our discrete diffusion model over probabilistic ones [5], [6] is that the exact solution can be found using mathematical programming. This enables us to study the exact behavior of the seeding size when the number of propagation hop varies.

We formulate the CFM problem as an $0-1$ Integer Linear Programming (ILP) problem below.

$$\text{minimize} \sum_{v \in V} x_v^0 \qquad \text{(VI.1)}$$

$$\text{subject to} \sum_{v \in V} x_v^d \geq \tau |V| \qquad \text{(VI.2)}$$

$$\sum_{w \in N(v)} x_w^{i-1} + \lceil \rho \cdot d(v) \rceil x_v^{i-1} \geq \lceil \rho \cdot d(v) \rceil \; x_v^i$$

$$\forall v \in V, i = 1..d \qquad \text{(VI.3)}$$

$$x_v^i \geq x_v^{i-1} \qquad \forall v \in V, i = 1..d \qquad \text{(VI.4)}$$

$$x_v^i \in \{0, 1\} \qquad \forall v \in V, i = 0..d \qquad \text{(VI.5)}$$

$$\text{where } x_v^i = \begin{cases} 0 & \text{if v is inactive at round } i \\ 1 & \text{otherwise} \end{cases}.$$

The objective of the ILP is to select a minimum number of seeds at the beginning. The constraint (VI.2) guarantees at least a fraction τ of nodes are activated at the end; the constraints (VI.3) capture the propagation model; and the constraint (VI.4) is simply to keep vertices active once they are activated.

We solve the ILP problem on Erdos collaboration networks, the social network of famous mathematician [23], with $\tau = 1$. The network consists of 6100 vertices and 15030 edges. The ILP is solved with the optimization package GUROBI 4.5 on Intel Xeon 2.93 Ghz PC and setting the time limit for the solver to be 2 days. The running time of the IP solver increases significantly when d increases. For $d = 1, 2,$ and 3, the solver return the optimal solutions. However, for $d = 4$, the solver cannot find the optimal solutions within the time limit and returns sub-optimal solutions with relative errors at most 15%.

The optimal (or sub-optimal) seeding sizes are shown in Figs. 3a, 3b, and 3c for $\rho = 0.4, 0.6$ and 0.8, respectively. VirAds provides close-to-optimal solutions and performs much better Max Degree. Especially, when $\rho = 0.8$ the VirAds's seeding is only different with the optimal solutions by one or two nodes. In addition, VirAds only takes fractions of a second to generate the solutions.

As proven in Section III, the seeding takes a constant fraction of nodes in the network. For Erdos Colloboration Network, the seeding consists of 3.8% to 7% the number of nodes in the networks. Further, the seeding can consist as high as 20% to 40% nodes in the network for larger social networks in next section.

Although the mathematical approach can provide accurate measurement on the optimal seeding size, it cannot be applied for larger networks. The rest of our experiments measures the quality and scalability of our proposed algorithm VirAds on a collection of large networks.

B. Large Social Networks

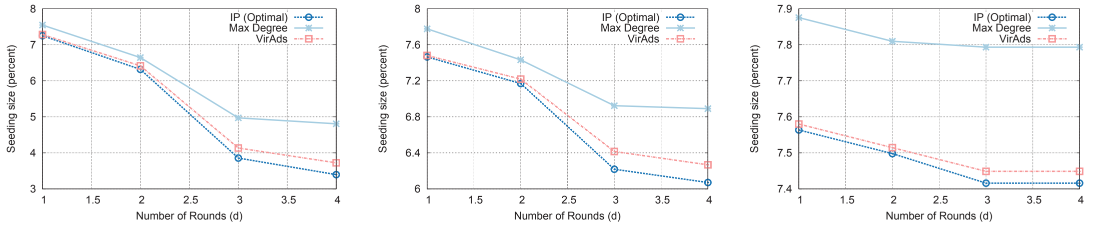
We select networks of various sizes including Coauthors network in Physics sections of the e-print arXiv [5], Facebook [31] and Orkut [32], a social networking run by Google. Links in all three networks are undirected and unweighted. The sizes of the networks are presented in Table I.

TABLE I
Sizes of the investigated networks

	Physics	**Facebook**	**Orkut**
Vertices	37,154	63,731	3,072,441
Edges	231,584	817, 090	223,534,301
Avg. Degree	12.5	25.6	145.5

Physics: We shall refer the physics coauthors network as Physics network or simply Physics. Each node in the network represents an author and there is an edge between two authors if they coauthor one or more papers. *Facebook* dataset consists 52% of the users in the New Orleans [31]. *Orkut* dataset is collected by performing crawling in last 2006 [32]. It contains about 11.3% of Orkut's users.

We compare our VirAds algorithm with the following heuristics *Random* method in which vertices are picked up randomly until forming a d-seeding and *Max Degree* method in which vertices with highest degree are selected until forming a d-hop seeding. Finally, we compare VirAds

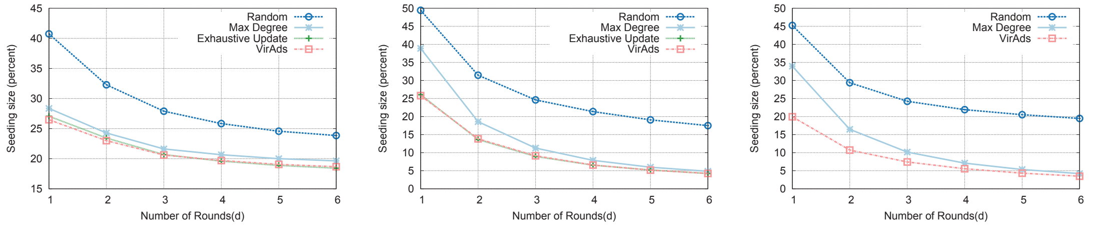


(a) $\rho = 0.4$

(b) $\rho = 0.6$

(c) $\rho = 0.8$

Fig. 3. Seeding size (in percent) on Erdos's Collaboration network. VirAds produces close to the optimal seeding in only fractions of a second (in comparison to 2 days running time of the IP(optimal))

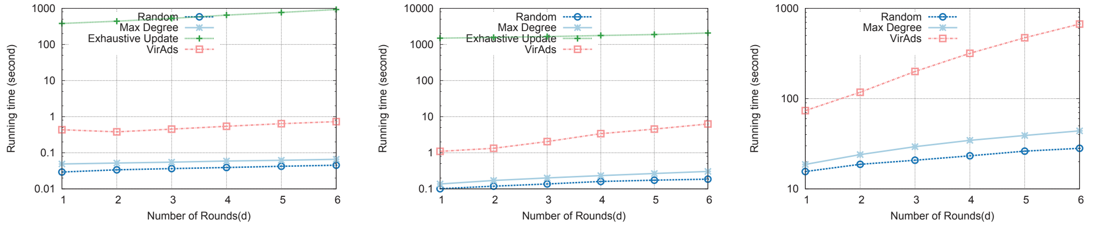


(a) Physics

(b) Facebook

(c) Orkut

Fig. 4. Seeding size when the number of propagation hop d varies ($\rho = 0.3$). VirAds consistently has the best performance.

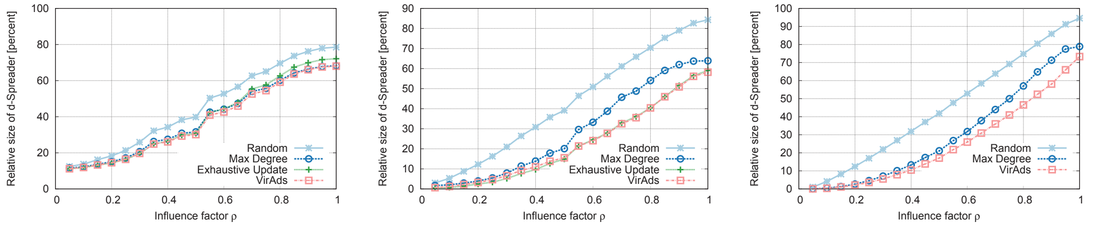


(a) Physics

(b) Facebook

(c) Orkut

Fig. 5. Running time when the number of propagation hop d varies ($\rho = 0.3$). Even for the largest network of 110 million edges, VirAds takes less than 12 minutes.



(a) Physics

(b) Facebook

(c) Orkut

Fig. 6. Seeding size at different influence factors ρ (the maximum number of propagation hops is $d = 4$).

with its naive implementation, called *Exhaustive Update*, in which after selecting a vertex into the seeding, the effectiveness of all the remaining vertices are recalculated. With more accurate estimation on vertex effectiveness, Exhaustive Search is expected to produce higher quality solutions than those of VirAds. To compare the performance of the presented algorithms, we vary the number of rounds d, and the influence factor ρ. The value of τ is set to one for this part.

Solution Quality. The seeding size with different number of propagation hop d when $\rho = 0.3$ are shown in Fig. 4. To our surprise, VirAds even performs equal or better than *Exhaustive Update* despite that it uses significantly less effort to update vertex effectiveness. VirAds has smaller seeding in Physics than *Exhaustive Update*; both of them give similar results for Faceboook; while *Exhaustive Update* cannot finish on Orkut after 48 hours and was forced to terminate. Sparingly update the vertices' effectiveness turns out to be efficient enough since the influence propagation is locally bounded. In addition, the VirAds's solutions are almost two times smaller than *Random*'s.

As shown in Fig. 4, the gap between VirAds and Max Degree is narrowed when the number of maximum hops increases. Hence, selecting nodes with high degrees as seeding is a good long-term strategy, but might not be efficient for fast propagation when the number of hops is limited. In Facebook and Orkut, when $d = 1$, *Max Degree* has 60% to 70% more vertices in the seeding than *VirAds*. In Physics, the gap between VirAds and the *Max Degree* is less impressive. Nevertheless, VirAds consistently produces the best solutions in all networks.

Scalability. The running time of all methods at different propagation hop d are presented in Fig 5. The time is measured in second and presented in the log scale. The running times increase slightly together with the number of propagation rounds d, and are proportional to the size of the network. The *Exhaustive Update* has the worst running time, taking up to 15 minutes for Physics, 20 minutes for Facebook. For Orkut, the algorithm cannot finish within 2 days, as mentioned. The three remaining algorithms *VirAds*, *Max Degree*, and *Random* take less than one second for Physics, and less than 10 seconds for Facebook. Even on the largest network Orkut with more than 220 million edges, VirAds finishes in less than 12 minutes.

Influence factor. We study the performance of VirAds and the other method at different influence factor ρ. The number of propagation rounds d is fixed to 4. The size of d-seeding sets are shown in Figures 6. VirAds is clearly still the best performer. The VirAds's solutions are nearly 5 times smaller than the solutions of Max Degree for small ρ (although it's hard to see this on the charts due to small seeding sizes).

Since all tested networks are social networks with small diameter, the seeding sizes go to zero when ρ is close to zero. The exception is the Physics, in which the seeding sizes do not go below 10% the number of vertices in the networks even when $\rho = 0.05$. A closer look into the Physics network reveals that the network contain many isolated cliques of small sizes (2, 3, 4, and so on) which correspond to authors that appear in only one paper. In each clique, regardless of the threshold ρ, at least one vertex must be selected, thus the seeding size cannot get below the number of isolated cliques in the networks. To eliminate the effect of isolated cliques, a possible approach is to restrict the problem to the largest component in the network.

C. Phases of Going Viral

In this part, we vary the value of the desired coverage τ, focusing on the case $\tau < 1$, to discover the effective seeding size. Specifically, we measure the *cost-effectiveness*, the ratio between the number of influenced nodes and the size of seeding every time a new node is selected (the higher cost-effectiveness, the better). The cost-effectiveness of seeding produced by VirAds and Maxdegree for $\rho = 0.2$ are shown in Figs. 7 and 8, respectively. When $d = 4$, one seed can activate around 100 other nodes at peak in Physics and Facebook, while the cost-effectiveness in Orkut is more than 250 at peak.

For $\rho = 0.2$ and other small values of ρ, a common trend in three networks is that the cost-effectiveness decreases when the seeding gets larger as shown in both Figs. 7 and 8. Moreover, we observe three different phases of the propagation process, which can be best seen in the Facebook network with $d = 4$. In the first phase, the influence propagate quickly into the network with high cost-effectiveness. The cost-effectiveness decreases quickly together when more nodes are added to the seeding. When a sufficient number of nodes in the network are influenced, the propagation process gains enough momentum to get into the second phase in which the cost-effectiveness increases sharply. The cost-effectiveness gets to a (local) peak again at the end of the second phase. Finally, in the third phase the cost effectiveness decays rapidly i.e. it costs much more to influence the last uninfluenced nodes in the network.

Although the cost-effectiveness peak at the beginning of the first phase, the number of influenced nodes is often too small at that time. Thus we argue that the best strategy is to *stop expanding the seeding before the third phase where we can achieve both the high cost-effectiveness and high coverage* τ i.e. the fraction of influenced nodes after d rounds. For example, at the end of the second phase we can influence around 8% of the nodes in the networks by selecting only around 0.1% of nodes in the seeding (Physics and Facebook).

As shown in Figs. 7 and 8, the "local peak" at the end of the second phase happens at the same time for different number of propagation rounds d. However, that "local peak" happens earlier with VirAds algorithm than Maxdegree. This gives another evidence that VirAds can select more effective seeding than Maxdegree.

Cost-effectiveness for $\rho = 0.8$ is shown in Fig. 9. For higher values of the influence factor ρ, only the first phase can be observed. The cost-effectiveness decreases quickly in all three networks. In addition, there is almost no difference among the cases $d = 2$, $d = 3$, and $d = 4$ which indicates that adding more propagation rounds does not help much, when the content has little incentive to go viral. This is consistent with the theory result developed in Section III. The same observation can be also made by looking at Figs. 10 and 11 in which we show the seeding size (in percent) to get a certain coverage τ in the network.

VII. Conclusions

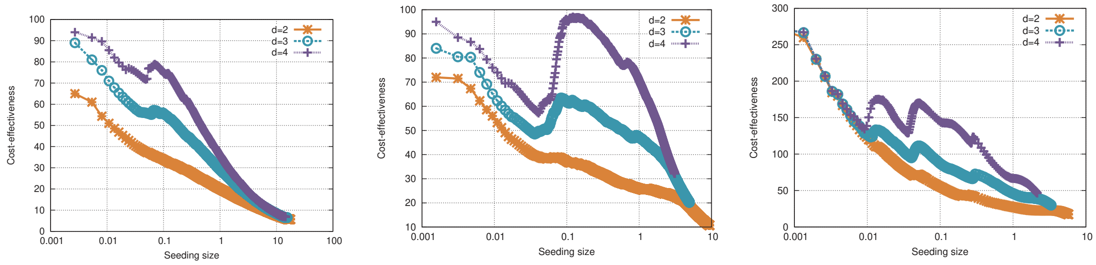
Massively advertising involves costly seeding when imposing the limit on the propagation. Moreover, allowing more time for the content to propagate does not help, if the content has little incentive of going viral. Thus, smarter viral marketing strategies which lead to actual sales are needed. For example, one of our future work is to design algorithms to 'direct' the influence to potential customers (e.g. women with respect to fashion products) rather than trying to spread the influence to every users in the networks.

References

[1] "Facebook statistics 2012," http://newsroom.fb.com/Key-Facts.

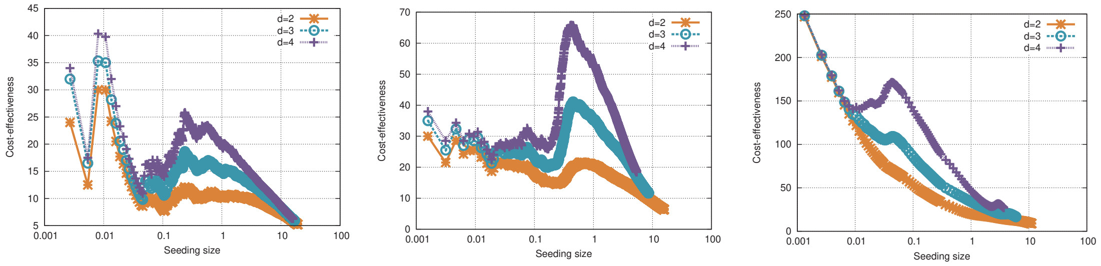
[2] "Alexa 2012," http://www.alexa.com/topsites.

[3] C. Shirky. (2011) The Political Power of Social Media: Technology, the Public Sphere, and Political Change. [Online]. Available: http://www.gpia.info/files/u1392/Shirky_Political_Poewr_of_Social_Media.pdf



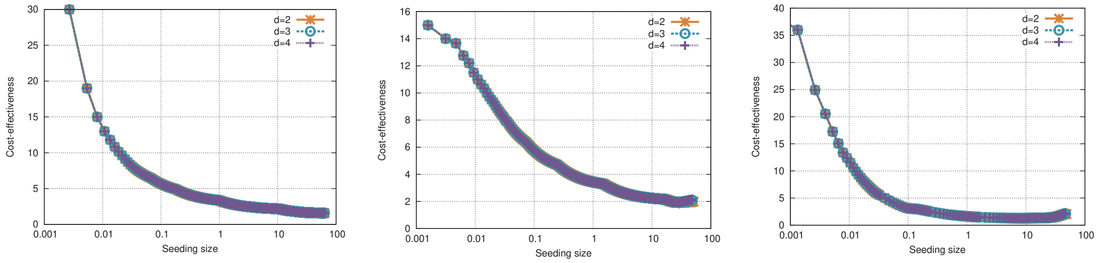
(a) Physics (b) Facebook (c) Orkut

Fig. 7. Cost-effectiveness of VirAds when $\rho = 0.2$.



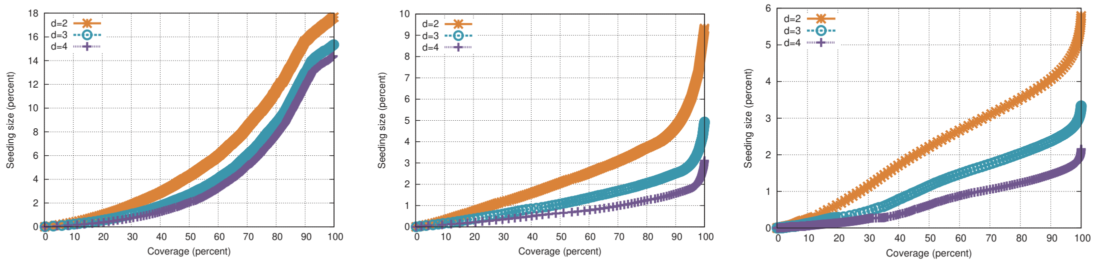
(a) Physics (b) Facebook (c) Orkut

Fig. 8. Cost-effectiveness of Maxdegree when $\rho = 0.2$.



(a) Physics (b) Facebook (c) Orkut

Fig. 9. Cost-effectiveness of VirAds when $\rho = 0.8$. Increasing the propagation round d does not boost the cost-effectiveness.



(a) Physics (b) Facebook (c) Orkut

Fig. 10. Seeding size of VirAds when $\rho = 0.2$.

[4] L. Rao, "Toyota Turns To Twitter To Repair Its Image," http://techcrunch.com/2010/03/02/toyota-turns-to-twitter-to-repair-its-image/, Mar. 2010.

[5] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *KDD'03*. ACM New York, NY, USA, 2003, pp. 137–146.

[6] D. Kempe, J. Kleinberg, and E. Tardos, "Influential nodes in a diffusion model for social networks," in *ICALP '05*, 2005, pp. 1127–1138.

[7] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *KDD '09*. New York, NY, USA: ACM, 2009, pp. 199–208.

[8] N. Chen, "On the approximability of influence in social networks," *SIAM Journal of Discrete Mathematics*, vol. 23, no. 3, pp. 1400–1415, 2009.

[9] Y. Li, B. Q. Zhao, and J. C. S. Lui, "On modeling product advertisement in large-scale online social networks," *Networking, IEEE/ACM Transactions on*, vol. PP, no. 99, p. 1, 2011.

[10] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *WWW '09*. New York, NY, USA: ACM, 2009, pp. 721–730.

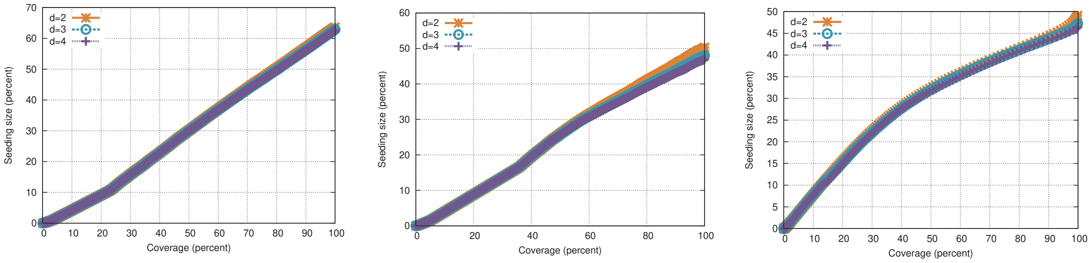
[11] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Trans. Web*, vol. 1, 2007.

[12] W. Aiello, F. Chung, and L. Lu, "A random graph model for power law graphs," *Experimental Math*, vol. 10, pp. 53–66, 2000.

[13] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *ACM KDD '07*. New York, NY, USA: ACM, 2007, pp. 420–429.

[14] M. Minoux, "Accelerated greedy algorithms for maximizing submodular set functions," in *Optimization Techniques*, ser. Lecture Notes in Control and Information Sciences, J. Stoer, Ed., 1978, vol. 7, pp. 234–243.

[15] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization



(a) Physics (b) Facebook (c) Orkut

Fig. 11. Seeding size of VirAds when $\rho = 0.8$. There are insignificant differences in seeding size when the number of propagation round d changes.

for prevalent viral marketing in large-scale social networks," in *ACM KDD '10*. New York, NY, USA: ACM, 2010, pp. 1029–1038.

[16] A. Goyal, F. Bonchi, L. Lakshmanan, and S. Venkatasubramanian, "On minimizing budget and time in influence propagation over social networks," *Social Network Analysis and Mining*, pp. 1–14, 2012.

[17] W. C., W. L., and N. Z., "Time-critical influence maximization in social networks with time-delayed diffusion process," in *AAAI*. AAAI Press, 2012.

[18] F. Wang, E. Camacho, and K. Xu, "Positive influence dominating set in online social networks," in *COCOA '09*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 313–321.

[19] F. Zou, Z. Zhang, and W. Wu, "Latency-bounded minimum influential node selection in social networks," in *WASA*, ser. LNCS, B. Liu, A. Bestavros, D.-Z. Du, and J. Wang, Eds., 2009, pp. 519–526.

[20] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Learning influence probabilities in social networks," *WSDM '10*, pp. 241–250, 2010.

[21] D. Peleg, "Local majority voting, small coalitions and controlling monopolies in graphs: A review," in *SIROCCO'96*, 1996, pp. 152–169.

[22] A. Barabasi, R. Albert, and H. Jeong, "Scale-free characteristics of random networks: the topology of the world-wide web," *Physica A*, vol. 281, 2000.

[23] A. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A: Statistical Mechanics and its Applications*, vol. 311, no. 3-4, pp. 590–614, 2002.

[24] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Reviews*, 2007.

[25] W. Aiello, F. Chung, and L. Lu, "A random graph model for massive graphs," in *STOC '00*. New York, NY, USA: ACM, 2000.

[26] ——, "Random evolution in massive graphs," *IEEE FOCS*, vol. 0, p. 510, 2001.

[27] A. Ferrante, "Hardness and approximation algorithms of some graph problems," 2006.

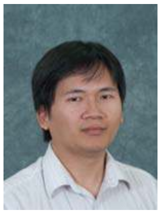
[28] T. N. Dinh, D. T. Nguyen, and M. T. Thai, "Cheap, easy, and massively effective viral marketing in social networks: truth or fiction?" in *Proceedings of the 23rd ACM conference on Hypertext and social media*, ser. HT '12. New York, NY, USA: ACM, 2012, pp. 165–174.

[29] U. Feige, "A threshold of ln n for approximating set cover," *Journal of ACM*, vol. 45, no. 4, pp. 634–652, 1998.

[30] T. N. Dinh, Y. Shen, D. T. Nguyen, and M. T. Thai, "On the approximability of positive influence dominating set in social networks," *Journal of Combinatorial Optimization*, pp. 1–17, 2012.

[31] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook," in *WOSN'09*, August 2009.

[32] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks," in *IMC'07*, San Diego, CA, October 2007.



Thang N. Dinh (S'11-M'14) received the Ph.D. degree in computer engineering from the University of Florida in 2013. He is an Assistant Professor at the Department of Computer Science, Virginia Commonwealth University. His research focuses on security and optimization challenges in complex systems, especially social networks, wireless and cyber-physical systems. He serves on TPC of several conferences including IEEE SOCIALCOM, SOCIALCOMNET, and was the publicity chair of CSoNet 2013. He is an Associate Editor of Computational Social Networks journal.

Huiyuan Zhang received her B.S. from Anhui University, Hefei, Anhui, China. Her area of interest is Networking Theory and Applied Mathematics. She has been a PhD student in CISE department, University of Florida since 2012, under the supervision of Dr. My T. Thai. Her area of interest is Networking Theory and Applied Mathematics.

Dzung T. Nguyen received the Ph.D. degree in computer engineering from University of Florida in 2013. His areas of interest are viral marketing on online social networks, vulnerability and cascading failures on coupled networks, and approximation algorithms for network optimization problems.

My T. Thai (M06) received the Ph.D. degree in Computer Science from the University of Minnesota, in 2005. She is an Associate Professor at the Computer and Information Science and Engineering Department, University of Florida. Her current research interests include algorithms and optimization on network science and engineering, with a focus on security. She has engaged in many professional activities, such as being the PC chair of EEE IWCMC 12, IEEE ISSPIT 12, and COCOON 2010. She is a founding EiC of Computational Social Networks journal, an Associate Editor of JOCO, IEEE Transactions on Parallel and Distributed Systems, and a series editor of Springer Briefs in Optimization. She has received many research awards including a UF Provosts Excellence Award for Assistant Professors, a DoD YIP, and an NSF CAREER Award.