Noname manuscript No.
(will be inserted by the editor)

The Unified Model of Social Influence and its Application in Influence Maximization

Ajitesh Srivastava · Charalampos Chelmis · Viktor K. Prasanna

Received: date / Accepted: date

Abstract The study of information dissemination on a social network has gained significant importance with the rise of social media. Since the true dynamics are hidden, various diffusion models have been exposed to explain the cascading behavior. Such models require extensive simulation for estimating the dissemination over time. In an earlier work, we proposed a Unified Model provides an approximate analytical solution to the problem of predicting probability of infection of every node in the network over time. Our model generalizes a large class of diffusion process. We demonstrate through extensive empirical evaluation that the error of approximation is small. We build upon our Unified Model develop an efficient method for influence maximization. Unlike most approaches, we assume that diffusion spreads not only via the edges of the underlying network, but also through temporal functions of external out-of-network processes. We empirically evaluate our approach and compare it against state of the art approaches on real-world large-scale networks. The evaluation demonstrates that our method has significant performance gains over widely used seed-set selection algorithms.

Keywords Analytical framework · Computational models · Diffusion models · Dynamical systems · Evolutionary models · Information cascades · Influence maximization

A. Srivastava
Department of Computer Science, University of Southern California, Los Angeles, USA
Tel.: +1213-740-9130
E-mail: ajiteshs@usc.edu

C. Chelmis, V. Prasanna
Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, USA

1 Introduction

With the proliferation of online social networks, researchers have tried to model, understand and make predictions of diffusion processes (Choi et al. 2010; Budak et al. 2012; Bakshy et al. 2012; Hajibagheri et al. 2013), such as the diffusion of innovations (Valente. 1996; Chelmis and Prasanna. 2013) and word-of-mouth recommendations (Kempe et al. 2003). Existing models of spreading processes in networks attempt to model diffusion as a result of social influence, i.e., the more influential a user is the wider the spread (Subbian et al. 2013). Diffusion is modeled using a network structure with static or dynamic edge probabilities (Kempe et al. 2003; Kleinberg. 2007), which are estimated from past observational data (Budak et al. 2012). According to such models, each node independently infects its neighbors with some probability, and each infected node then propagates the infection in the network. Even though this process captures individual influence (i.e., node-to-node), it ignores social influence effects which appear as a result of neighborhood or global pressure (Chelmis and Prasanna. 2013; Chelmis et al. 2014). (Subbian et al. 2013) proposed to mediate this problem by incorporating the notion of social capital to characterize the network effect in the influence process, whereas (Chelmis and Prasanna. 2013; Chelmis et al. 2014) presented agent-based computational models, which quantified pairwise influence and global dynamics in the spread of technology adoption at the workplace.

Two of the most widely used diffusion models are the Linear Threshold Model (LTM) (Granovetter. 1978), and the Independent Cascade Model (ICM) (Kempe et al. 2003). LTM assumes that a node gets infected when the number of its infected neighbors exceeds a threshold. According to ICM, each node has n independent chances to become infected; n being the number of its infected neighbors. ICM is closely related to the Susceptible - Infected - Susceptible

(SIS) and Susceptible - Infected - Recovered (SIR) models (Jacquez and Simon. 1993; Hethcote. 2000; Kamp. 2010). Furthermore, it was recently shown that ICM and LTM are special cases of the Genetic Algorithm Diffusion Model (GADM) (Lahiri and Cebrian. 2010). GADM emulates social interactions through a tail-swap cross-over interaction (Hajibagheri et al. 2013), assuming that social interactions are always pairwise. In our work, we propose a model to capture not only pairwise influence, but also local neighborhood effects, aggregate social behavior, and external factors, or a combination of the above.

Current models of influence and diffusion require expensive Monte Carlo simulations. Instead, we devise a novel formulation of progressive diffusion with minimum computational complexity. We provide a generalized, analytical solution to the diffusion mechanism that comprises of two processes unfolding over the network simultaneously: (a) pairwise influence, and (b) pressure from collective dynamics, which can be a result of local social pressure, global influence, or external forces, or a combination of the above. Our methodology is vertex-centric, i.e., models each user separately, offering great flexibility in terms of modeling personalized influence functions, and allows for the use of time-dependent influence functions. In a separate work (Srivastava et al. 2015 (Accepted), we have shown that this vertex-centric approach results in efficient parallel implementation with considerable speed-ups. Note that in this work, we are not concerned with learning the parameters that drive the spread of infection from observational data. While accurately modeling the underlying mechanism of infection is important, it is outside of the scope of this paper. To the best of our knowledge, our work is the first to (a) enable analytical computation of complex, non-linear phenomena like influence, while (b) considering multiple factors that can change over time, without requiring extensive simulation runs to estimate the propagation probabilities at the steady state. Our formula explicitly and formally unites a rich class of popular diffusion processes in social networks (Valente. 1996; Kempe et al. 2003; Kleinberg. 2007; Budak et al. 2012; Chelmis and Prasanna. 2013) as special cases.

We show the utility of our Unified Model in the task of Influence Maximization (Kempe et al. 2003), which poses the problem of finding k vertices that maximizes the spread of influence in a directed network under an infection model M. This problem is primarily motivated by applications to viral marketing, where the goal is to select individuals to target as part of a marketing intervention strategy in order to maximize subsequent cascade of product adoption (Kempe et al. 2003). Other applications include finding inoculation targets in epidemiology (Newman. 2002; Fan et al. 2014), opinion maximization (Gionis et al. 2013; Kelman. 1961) and efficient gateway-finding in large-scale graphs (Tong et al. 2010). The problem is known to be NP-hard, so approximation algorithms must be used (Kempe et al. 2003; Shakarian et al. 2013; Xu et al. 2014). Unlike most present approaches for influence maximization in networks, we assume that diffusion spreads not only from node to node via the edges of the underlying network like an epidemic, but also through influence of external out-of-network processes such as mass media (e.g., newspapers, TV stations and online news sites). To address the problem of influence maximization we propose a greedy solution based on our Unified Model, which we empirically evaluate against state of the art approaches on real-world large-scale networks. Our experiments demonstrate the efficiency and practical utility of our proposed algorithm.

The rest of the paper is organized as follows: Section 2 provides an overview of the most relevant related work which has been undertaken in this area. We formally describe our unified influence model in Section 3. We show how our model can be reduced into popular models for influence propagation in Section 4. We verify our model's generality by comparing it against extensive simulation results of other diffusion models using a real life dataset and a series of synthetic data. We present our quantitative evaluation results in Section 5. In Section 6 we propose a greedy seed set selection algorithm to maximize influence in social networks and evaluate its performance against state of the art influence maximization approaches. Finally, we discuss the findings of our work and draw our conclusions in Section 7.

2 Related Work

In prior work, we presented a generalized, analytical model of influence in social networks (Srivastava et al. 2014). We argued that our closed-form expression for the probability of infection for every node in an arbitrary, directed network at any time t captures social influence at various levels of granularity, ranging from pairwise influence, to local neighborhood, to the general population, and external events, therefore capturing the complex dynamics of human behavior. The present paper expands on our prior work both in breadth and in depth. Particularly, we extend the analysis of our unified model of influence in both real-world large scale social network datasets as well as synthetic random networks. We closely study the estimates produced by our analytical solution and compare them against several popular diffusion models. Subsequently, we perform a thorough and sound analysis of our unified model with the purpose of characterizing the quality of our approximation as a function of network properties and model parameters, and we empirically confirm that the approximation error of our solution is small and invariant of the topological and statistical properties of the undelying network. We apply our unified model to the problem of seed set selection with the goal of maximizing influence in social networks. We propose a greedy

solution based on our analytical solution and demonstrate significant improvements in the total number of infections with increasing seed-set sizes.

Information dissemination in online social networks has been thoroughly studied (Abrahamson and Rosenkopf. 1997; Leskovec et al. 2006; Bakshy et al. 2012; Kamp. 2010; Yang and Leskovec. 2010; Budak et al. 2012). Epidemic models (Hethcote. 2000) and computational approaches, such as threshold models (Granovetter. 1978), deterministic or stochastic models (Jacquez and Simon. 1993; Hajibagheri et al. 2013), and genetic algorithms (Lahiri and Cebrian. 2010) have been proposed. Out of the most widely used diffusion models, the General Threshold Model (Kempe et al. 2003), the Linear Threshold Model (Granovetter. 1978), and the Independent Cascade Model (Kempe et al. 2003), have been shown to be equivalent (Kempe et al. 2003). Independent Cascade Model is also closely related to the Susceptible-Infected-Susceptible and Susceptible-Infected-Removed models (Jacquez and Simon. 1993; Hethcote. 2000; Kamp. 2010).

Nearly all prior work attempts to model the diffusion process over a network structure, with static or dynamic edge probabilities, which are estimated from past observations. Social influence models to quantify users' influence in a social network have been proposed (Anagnostopoulos et al. 2008; Myers et al. 2012; Subbian et al. 2013). The role of network structure in information dissemination, and diffusion as a result of social influence was studied in (Bakshy et al. 2012). (Tang et al. 2009) proposed a model of topical affinity propagation to measure influence at the topic level on large networks. Such prior works assume that influence probabilities are given. (Goyal et al. 2010) proposed models to capture and learn influence in online social networks, whereas (Gomez Rodriguez et al. 2010) developed an approximation algorithm for inferring networks of diffusion and influence. External influence in networks has been considered in (Myers et al. 2012), but in that case all nodes share the same probability of external influence. Furthermore, the focus of that study was to infer the number of exposures incurred by external sources and learn the exposure curve. In this regard, our work introduces an important dimension to the diffusion process, which in our case explicitly encompasses pairwise influence, local neighborhood effects, aggregate social behavior and external factors.

It has been shown that exact computation of infection probabilities is #P-hard (Chen et al. 2010a). (Du et al. 2013) proposed a randomized algorithm for influence estimation in continuous-time diffusion networks. (Bóta et al. 2013) proposed several heuristics to calculate the a posteriori infection probabilities for all nodes in a graph for which all edge infection probabilities are given. Their work fits the Independent Cascade model, which we show to be a special case of our framework.

For the task of Influence Maximization an approximation algorithm was proposed by Kempe et. al. (Kempe et al. 2003) which is provably within 63% of the optimal for ICM and GLT. However, it is not scalable for large networks. Degree Discount heuristic (Chen et al. 2010a) has been shown to perform equally good on ICM, with much less computational complexity. CELF (Leskovec et al. 2007) and CELF++ (Goyal et al. 2011a) speed up the seed-set selection algorithm of Kempe et. al. by utilizing the sub-modularity property of influence models. For LTM, LDAG has been proposed (Chen et al. 2010b) that finds a local directed acyclic graph, claiming that most of the influence spreads in a local neighborhood. SPS-CELF++ (Goyal et al. 2011b) performs some further optimizations on CELF++, specifically for LTM. We demonstrate that such methods are not versatile enough to be applied on other influence models, by showing that our algorithm outperforms both LDAG and SPS-CELF++ for Linear Friendship Model (Budak et al. 2012). (Xu et al. 2014) describes a model that exploits the property of transitivity in influence which is prevalent in social networks. They achieve 0.8 approximation on optimal seed-set selection by treating Influence Maximization as a weighted maximum cut problem. However, their algorithm is not scalable; it takes $18,600s$ on a graph of $15,347$ nodes. In contrast, our algorithm runs in less than a minute for a similar sized graph.

3 Analytical Model of Influence

In this section, we provide the description and mathematical formulation of our proposed unification model of influence in social networks. We model the social network as a directed graph $G = (V, E)$, where a node $v \in V$ represents an individual, and edge $(v, u) \in E$ exists if v interacts with u (in our context v influences u). For every node v, we define the set of incoming neighbors $N_i(v) = \{u|(u, v) \in E\}$, and the set of outgoing neighbors $N_o(v) = \{u|(v, u) \in E\}$. Our goal is to model the probability of infection for every node in the network at any time t. Typically, in a diffusion process, a node can exist in one of two states at a given time - infected or susceptible. Here, we study the problem of progressive diffusion, where nodes that are infected do not become healthy again, i.e., they do not return to the susceptible state. Hence, every node can be infected once, and once infected stays infected. Healthy nodes cannot infect others.

3.1 Unified Model of Influence

We start with a seed set $S \subset V$ of infected nodes at time $t = 0$. The infection process proceeds in discrete time steps, in which two types of influence unfold over the network

(Chelmis and Prasanna. 2013). According to the first process, each infected node v attempts to infect its neighbors (*individual* influence). Each attempt of infecting node $u \in N_o(v)$ has a chance of success, but the probability of infection $p_{(v,u)}(t)$ is pairwise and may change over time. Note that we assume independence between infection attempts from multiple neighbors. The second source of influence we consider is *collective* influence. According to this process, each susceptible node u can be infected with probability $r_u(t)$, independent of individual influence. This may include external factors (Abrahamson and Rosenkopf. 1997; Bass. 2004; Choi et al. 2010), or external sources of exposure (Myers et al. 2012), or the status of the incoming neighborhood of u (Chelmis and Prasanna. 2013). A couple of things should be noted here. First, function $r_u(t)$ is node specific, and may be time dependent. Second, there may be arbitrary number of collective influence attempts on a node, as we assume $r_u(t)$ is not conditioned upon the node already having undergone a collective influence attempt or not. The process repeats until a pre-specified stopping criterion is satisfied (e.g., number of time steps elapsed, or fraction of infected nodes has exceeded some number).

3.2 Infection Probability Formula Under the Unified Model

Let $B_{u,t}$ represent the probability of infection of node u by the time t. Initial values $\{B_{u,0}\}$ are either 0 or 1 depending on the membership of u in the seed set. Let $E_{v,t}$ denote the indicator variable, which is 1 if node v is infected by the time t, 0 otherwise. To find the probability of a node u being infected at time t, we consider an arbitrary ordering of its incoming neighbor set $N_i(u)$: $< v_1, v_2, \ldots, v_n >$. Based on this, we define *zero state probability* at time $t-1$: $P^0_{s_n,s_{n-1},\ldots,s_1}$, where superscript 0 denotes $E_{u,t-1} = 0$. The subscript is a vector, which elements s_i denote the value of $E_{v_i,t-1}$, and can take values in $\{0,1,*\}$. $s_i = 0$ represents $E_{v_i,t-1} = 0$, $s_i = 1$ denotes $E_{v_i,t-1} = 1$, and $s_i = *$ indicates marginalization over the state of v_i, i.e., '$E_{v_i,t-1} = 0$ or 1'. For instance, for a node u with four neighbors, $P^0_{0,1,*,1}$ denotes the probability $P(E_{u,t-1} = 0, E_{v_4,t-1} = 0, E_{v_3,t-1} = 1, E_{v_1,t-1} = 1)$. We begin by calculating $B_{u,t}$ in the special case of G being a tree, i.e., each node has at most one incoming neighbor.

Corollary 1 *The infection probability of node u with parent v in a tree is given by:*

$$B_{u,t} = 1 - (1 - r_u(t))\big((1 - p_{v,u}(t))(1 - B_{u,t-1}) + p_{v,u}(t)(1 - B_{v,t-1})\prod_{k=1}^{t-1}(1 - r_u(k))\big). \quad (1)$$

Proof The probability of node u not being infected by time t is $P(E_{u,t} = 0) = 1 - B_{u,t}$. Either one of two things must have happened for u not to be infected by time t. First, state $E_{u,t} = 0$ was reached from state $(E_{v,t-1} = 0, E_{u,t-1} = 0)$ if and only if collective influence $r_u(t)$ failed to infect u at time t. Intuitively, when the parent of u was not infected at time$t-1$, the only chance for u to be infected at time t is through collective influence $r_u(t)$, with probability $1 - r_u(t)$. Second, state $E_{u,t} = 0$ was reached from state $(E_{v,t-1} = 1, E_{u,t-1} = 0)$, i.e., when the parent of u was infected at time $t-1$, if and only if collective influence was unsuccessful, and furthermore v failed to infect u. As the two processes are independent, this can happen with probability $(1 - r_u(t))(1 - p_{v,u}(t))$. It follows that,

$$\begin{aligned} 1 - B_{u,t} &= P^0_1(1 - r_u(t))(1 - p_{v,u}(t)) + P^0_0(1 - r_u(t)) \\ &= (1 - r_u(t))(P^0_1(1 - p_{v,u}(t)) + P^0_0) \\ &= (1 - r_u(t))((P^0_* - P^0_0)(1 - p_{v,u}(t)) + P^0_0) \\ &= (1 - r_u(t))(P^0_*(1 - p_{v,u}(t)) + P^0_0 p_{v,u}(t))), \end{aligned} \quad (2)$$

where $P^0_* = P(E_{u,t-1} = 0) = 1 - B_{u,t-1}$ and $P^0_0 = P(E_{u,t-1} = 0, E_{v,t-1} = 0)$. This means v and u were both susceptible at time $t-1$. If v was also not infected by the time $t-1$, u can only be susceptible because all collective influence till that time failed, i.e., $P^0_0 = (1 - B_{v,t-1})\prod_{k=0}^{t-1}(1 - r_u(k))$. We set $r_u(0) = 1$ if $u \in S$, 0 otherwise. Substituting the values of P^0_* and P^0_0 in Equation 2, results in Equation 1.

Next, we extend Equation 1 to a graph of any type. Without loss of generality, we focus on directed graphs, as undirected graphs can be converted into their directed equivalent.

Lemma 1 *The probability of a node u not being infected by the time t is related to the zero state probabilities as follows*

$$1 - B_{u,t} = (1 - r_u(t)) \sum_{s_i \in \{0,*\}} \left(P^0_{s_n,s_{n-1},\ldots,s_1} \prod_{i=1}^{n} (1 - p_{v_i,u}(t))^{\delta_{s_i,*}} \prod_{i=1}^{n} p_{v_i,u}(t)^{\delta_{s_i,0}} \right). \quad (3)$$

where $\delta_{a,b} = 1$ only if $a = b$, 0 otherwise, is the Kronecker delta function.

Proof When the number of incoming neighbors is one, Lemma 1 follows from Equation 2. Now, suppose the statement is true for $k \geq 1$ parents. Consider a sequence $\mathbf{x_k} = < s_k, s_{k-1}, \ldots, s_1 >$. We look at the new terms that are added due to the inclusion of v_{k+1}. For ease of notation, let $D(\mathbf{x_k}) = (1 - r_u(t))\prod_{i=1}^{k}(1 - p_{v_i,u}(t))^{\delta_{s_i,*}} \prod_{i=1}^{k} p_{v_i,u}(t)^{\delta_{s_i,0}}$. Equation 3 can be rewritten as

$$1 - B_{u,t} = \sum_{\mathbf{x_n}} P^0_{\mathbf{x_n}} D(\mathbf{x_n}). \quad (4)$$

We have assumed that this is true for $n = k$. The addition of v_{k+1} affects $P(E_{u,t})$ in ways similar to those discussed in Corollary 1, i.e., if $E_{v_{k+1},t-1} = 1$, then this new node fails to infect u with probability $(1 - p_{v_{k+1},u}(t))$. On the other hand,

if $E_{v_{k+1},t-1} = 0$, node v_{k+1} does not have the ability to infect. Formally, the new terms added are:

$$\begin{aligned}
&P^0_{1,\mathbf{x_k}}(1-p_{v_{k+1},u}(t))D(\mathbf{x_k}) + P^0_{0,\mathbf{x_k}}D(\mathbf{x_k}) \\
=&(P^0_{*,\mathbf{x_k}} - P^0_{0,\mathbf{x_k}})(1-p_{v_{k+1},u}(t))D(\mathbf{x_k}) + P^0_{0,\mathbf{x_k}}D(\mathbf{x_k}) \\
=&P^0_{*,\mathbf{x_k}}(1-p_{v_{k+1},u}(t))D(\mathbf{x_k}) + P^0_{0,\mathbf{x_k}}p_{v_{k+1},u}(t)D(\mathbf{x_k}) \\
=&P^0_{*,\mathbf{x_k}}D(*,\mathbf{x_k}) + P^0_{0,\mathbf{x_k}}D(0,\mathbf{x_k}),
\end{aligned}$$

which would generate the required terms in the right hand side of Equation 4, when $n = k+1$. This indicates that the statement is true for $k+1$ incoming neighbors. By induction, Lemma 1 is true $\forall n$.

Theorem 1 *An approximate probability of infection is given by the recurrence relation:*

$$\begin{aligned}
B_{u,t} = 1 - \Bigg[&(1-B_{u,t-1})\Bigg(\prod_{v\in N_i(u)}(1-p_{v,u}(t)B_{v,t-1})\Bigg) \\
&+\Bigg(\prod_{v\in N_i(u)}p_{v,u}(t)(1-B_{v,t-1})\Bigg) \\
&\Bigg(\prod_{k=1}^{t-1}(1-r_u(k)) - 1 + B_{u,t-1}\Bigg)\Bigg](1-r_u(t)).
\end{aligned} \tag{5}$$

The approximation comes from assuming that the states of infection of incoming neighbors of a given node u are independent, i.e., for two incoming neighbors v_i and v_j, events $E_{v_i,t-1} = 0$ and $E_{v_j,t-1} = 0$ are independent. Next, we proceed with proving the Theorem.

Proof We attempt to find the zero state probabilities for sequence $\mathbf{x_n}$. When $\mathbf{x_n} =< 0,0,\ldots,0 >$, u and all nodes in $N_i(u)$ are susceptible, which means that collective influence till $t-1$ was unsuccessful. Further, at $k = 0$, $r_u(t) = 1$ for $u \in S$. In this case,

$$P^0_{0,0,\ldots,0} = \prod_{k=0}^{t-1}(1-r_u(k)) \prod_{v\in N_i(u)}(1-B_{v,t-1}). \tag{6}$$

Any other sequence $\mathbf{x_n}$, which consists of at least one $*$ in the i-th position, represents the state of u being not infected by the state of its i-th neighbor. Given the state of u's neighbors, the conditional probability of u not being infected is $1-B_{u,t-1}$. The zero state probability is then computed as follows

$$P^0_{\mathbf{x_n}} = (1-B_{u,t-1}) \prod_{s_i=0}(1-B_{v_i,t-1}). \tag{7}$$

Combining Equations 4, 6 and 7, results in the following:

$$\begin{aligned}
\frac{1-B_{u,t}}{1-r_u(t)} = &(1-B_{u,t-1})\Bigg(\prod_{v_j\in N_i(u)}(1-p_{v_j,u}(t))\Bigg) \\
&\Bigg(\sum_{\mathbf{x_n}}\prod_{s_j=0}\frac{(1-B_{v_j,t-1})p_{v_j,u}(t)}{1-p_{v_j,u}(t)}\Bigg) \\
+&\Bigg(B_{u,t-1} - 1 + \prod_k(1-r_u(t))\Bigg)\Bigg(\prod_{v_j}(1-B_{v_j,t-1})p_{v_j,u}(t)\Bigg).
\end{aligned}$$

After simplification, the above equation reduces to Equation 5. This step completes the proof.

3.3 Complexity Analysis

The recurrence relation in Equation 5 requires inspection of all incoming links to a node u, $|N_i(u)|$, at every time step. Therefore, in order to evaluate Equation 5 for all nodes for t time steps, the number of operations required is $t\sum_u O(|N_i(u)|) = O(|E|t)$. In contrast, a discrete-time influence model that relies on Monte-Carlo simulations to obtain the expected infection probability at time t would require $O(R|E|t)$ operations, where R is the number of simulations.

4 Reduction to Other Models

Our analytical formula of influence in social networks, offers great flexibility in terms of modeling a variety of diffusion processes. Specifically, popular diffusion models can be reduced to special cases of the Unified Model, by carefully defining the individual influence probabilities and collective influence functions. We next describe few such reductions.

4.1 Complex Contagion Model

According to the Complex Contagion Model (Chelmis and Prasanna. 2013), infection can be achieved at time t in two ways. First, each node that was infected at time $t-1$ attempts to infect each of its outgoing neighbors with probability p. Once a node is infected, it cannot be infected again. Once all infected nodes are examined, healthy nodes have a chance of random infection based on the popularity of the contagion at time $t-1$. Particularly, for n_i^{t-1} infected nodes by the time $t-1$, the probability of random infection at time t is given by an exponential growth law: $r(t) = \exp(\alpha n_i^{t-1} - \beta)$, where α and β are constants (Chelmis and Prasanna. 2013).

Proposition 1 *The Complex Contagion Model (Chelmis and Prasanna. 2013) is a special case of the Unified Model (Section 3.1), and hence it can be approximated by Equation 5,*

when pairwise individual influence is constant and time independent, and collective influence is equivalent to random infection.

Proof (Reduction) We begin with Equation 5. We model individual influence as

$$p_{v,u}(t) = p, \forall v,u,t\,. \tag{8}$$

Substituting collective influence with the random infection factor results in

$$r_u(t) = r(t) = \exp(\alpha \sum_u B_{u,t-1} - \beta)\,, \tag{9}$$

since the number of infections by the time $t-1$, n_i^{t-1}, is computed as follows:

$$n_i^{t-1} = \mathbb{E}(\sum_u E_{u,t-1}) = \sum_u \mathbb{E}(E_{u,t-1}) = \sum_u B_{u,t-1}\,. \tag{10}$$

Equations 8 and 9 form the reduction of Unified Model to the Complex Contagion Model.

4.2 Independent Cascade Model

In the Independent Cascade Model (Kempe et al. 2003), a seed set of infected nodes is provided. At each time step t, each node is either infected or susceptible, and every node v that was infected at time $t-1$ has a single chance to infect each of its neighbors u. The infection succeeds with probability $p_{v,u} = p$.

Proposition 2 *The Independent Cascade Model is a special case of the Unified Model (Section 3.1), when collective influence is a function of the state of infection of nodes in the local neighborhood.*

Proof (Reduction) We begin with Equation 5. At any time t, a susceptible node u has a single chance to be infected by its neighbors that were infected at $t-1$. If at least one of them succeeds, u gets infected. The probability of node u getting infected is then given by

$$\begin{aligned} r_u(t) &= P(\text{at least one infected neighbor succeeds}) \\ &= 1 - P(\text{no infected neighbor succeeds}) \\ &= 1 - \prod_{v \in N_i(u)} (1 - p(A_{v,t-1}))\,, \end{aligned} \tag{11}$$

where $A_{v,t-1}$ is the probability of v being infected at time $t-1$. Since $B_v^t = \sum_{\tau=0}^{t} A_{v,\tau}$, if follows that:

$$A_{v,t} = \begin{cases} B_{v,t-1} & \text{if } t = 1 \\ B_{v,t-1} - B_{v,t-2} & \text{if } t > 1 \end{cases}$$

This step concludes the reduction.

4.3 Threshold Models

In threshold models the probability of infection of a node depends on the popularity of the contagion in its incoming neighborhood. Several threshold models exist in the literature, including the Linear Threshold Model (Granovetter. 1978), and the Linear Friendship Model (Anagnostopoulos et al. 2008; Budak et al. 2012). The Generalized Threshold Model (Kempe et al. 2003) dictates that a node u is infected based on a threshold $\theta_u \in [0,1]$ and a monotone function of the set of its infected neighbors $f(In(u,t)) \in [0,1]$. Particularly, u is infected at time t if $f(In(u,t)) \geq \theta_u$. Note that the the threshold θ_u can be randomly selected at each time t (Lu et al. 2012) leading to non-determinism of the infection process. Since, these thresholds are selected uniformly at random, this is equivalent to saying that the probability of infection of a healthy node u at time t is $f(In(u,t))$.[1]

Proposition 3 *The Generalized Threshold Model is a special case of the Unified Model (Section 3.1), when pairwise individual influence is zero, and collective influence is a function of weighted influence from the local neighborhood of nodes.*

Proof (Reduction) We begin with Equation 5. At any time t, the probability of node u getting infected is given by a function of u's status as follows:

$$P(u \text{ infected at time } t) = f(In(u,t), \mathbf{b_u})\,, \tag{12}$$

where $In(u,t) = \{v | v \in N_i(u), E_{v,t-1} = 1\}$ and $\mathbf{b_u} = \{b_{v,u} | v \in N_i(u)\}$ is a vector of pairwise weights $b_{v,u}$ associated to v's incoming neighbors. Substituting $r_u(t)$ in Equation 5 with Equation 12, and setting all individual influence probabilities to zero, concludes the reduction.

Linear Friendship Model: The Linear Friendship Model (LFM) (Anagnostopoulos et al. 2008; Budak et al. 2012) models the additive effect to the probability of infection at time t as a linear function of infected neighbors by the time $t-1$, and applies logistic regression to fit the linear function into a probability value.The Linear Friendship Model (Anagnostopoulos et al. 2008; Budak et al. 2012) can be treated as a special case of the Unified Model (Section 3.1). The reduction follows similar reasoning to that for Generalized Threshold Model. The difference lies in the function used to model the effect of the local neighborhood to a node's probability of infection in Equation 12. Here, the probability of node u getting infected at time t is given by

$$P(u \text{ infected at time } t) = \frac{\exp(\alpha |In(u,t)| + \beta)}{1 + \exp(\alpha |In(u,t)| + \beta)}, \tag{13}$$

[1] Note that this is different from the linear threshold model in (Kempe et al. 2003) in that the threshold may change at every time step.

where $|In(u,t)| = \sum_{v \in N_i(u)} B_{v,t-1}$, i.e., the number of infections by the time $t-1$ is calculated similarly to the Complex Contagion Model (Chelmis and Prasanna. 2013) using Equation 10, with the difference that the population is restricted to the local neighborhood of node u. Including pairwise weights $b_{v,u}$ into the formulation results in $|In(u,t,\mathbf{b_u})| = \sum_{v \in N_i(u)} b_{v,u} B_{v,t-1}$, which concludes the reduction.

5 Experiments

With our model well-defined, we now apply it to a real life dataset from popular social news aggregator Digg[2], and a series of synthetic data. First, to better illustrate the ability of our Unified Model (Theorem 1 in Section 4) to capture real-life behavior, we examine a specific real-world case study where we estimate information diffusion in a dynamic social network. We compare the results of our analytical framework, with those produced by several popular diffusion models (CCM, LT, LFM, and ICM in Section 4). Particularly, we verify that the expected epidemics calculated using Theorem 1 matches very well the average outcome of multiple simulation runs of these models. Subsequently, we run a series of large-scale experiments on synthetic data to show that the approximation error is small and insensitive both to graph properties and to models' parameters. Note, that in this work, we are not concerned with learning the spread of infection from observational data. While this aspect is important, it is outside of the scope of this paper. Our findings imply that Equation 5 is able to accurately predict the expected epidemics forecasted by the rest of the models without extensive numerical simulations.

5.1 Experiments Using Real-World Data

We used a subset of Digg's[3] follower graph (Lin et al. 2009). Digg is a popular social news aggregator that allows users to collectively curate a list of news stories they find online by submitting them to Digg and voting for them. In addition, Digg allows users to form social networks by designating as friends users whose activities they would like to track. Our dataset consists of $1,244$ nodes and $28,343$ directed links. A link from v to u exists if v influences u, i.e., when u follows v. Table 1 summarizes the set of parameters used in our experiments. For each model, we start with a seed set of two infected nodes.

Figure 1 shows the results of infection spreads over time using average of 1000 simulations for each model, and the corresponding predicted values obtained by using the analytical solution from Theorem 1. The prediction matches very well with the average simulations, providing an empirical, quantitative confirmation that Equation 5 produces a good fit to the expected outcome which is obtained by computationally expensive simulation runs.

Table 1 Parameters used in the experimental validation on Digg follower graph

parameter set 1	CCM	$p = 0.1, r(t) = exp(0.002 n_i^{t-1} - 6)$
	GLT	$f(In(u,t), \mathbf{b_u}) = \sum_{v \in In(u,t)} b_{v,u}$
	ICM	$p = 0.1$
parameter set 2	CCM	$p = 0.01, r(t) = exp(0.002 n_i^{t-1} - 6)$
	GLT	$f(In(u,t), \mathbf{b_u}) = \frac{\exp(\lvert In(u,t,\mathbf{b_u})\rvert)}{1+\exp(\lvert In(u,t,\mathbf{b_u})\rvert)}$
	ICM	$p = 0.7$

5.2 Experiments on Erdős-Rényi Random Graphs

For an extensive analysis of the approximation error, we run a series of experiments on simulated data. To study the effect of graph size on the approximation quality, we generated random sparse directed graphs of sizes 20, 40, 80, 160, 320, 640, $1,280$, $2,560$, $5,120$ and $10,240$ following the Erdős-Rényi model (Erdős and Rényi. 1960), with number of edges approximately five times the number of nodes. For each size and a given set of parameters we generated a random graph, nodes of which were uniformly partitioned into five roughly equal cardinality subsets. We started with infecting all nodes in one of these subsets, and ran 1000 simulations. Thus, we have five initial conditions for each size and set of parameters for a given model, and for each of the initial conditions we ran 1000 simulations. To examine the effect of graph density[4] on the approximation error, we fixed the size of the graph and then we generated random directed graphs with varying density values of $0.002, 0.004, 0.008, ..., 0.512$. We repeated our experiments for graphs with fixed size, but varied density instead. The parameters used for each model are summarized in Table 2.

We report approximation error using two measures: (a) root mean squared error (RMSE) at time t, and (b) fractional error in prediction of total number of infections at time t. We measure the error in approximating the probability of infection at time t in terms of RMSE as follows:

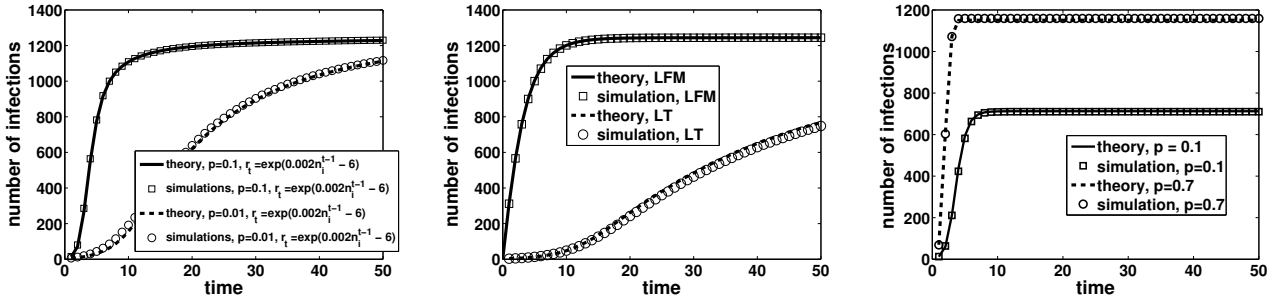
$$e_t^{rms} = \sqrt{\frac{\sum_u (B^*_{u,t} - B_{u,t})^2}{n}}, \tag{14}$$

where $B_{u,t}$ is the probability of infection of node u by the time t obtained by simulations, and $B^*_{u,t}$ is the value predicted by Theorem 1. We further report the fractional error

[2] `http://digg.com/`

[3] The dataset can be found online at `http://www-scf.usc.edu/~ajiteshs/datasets/digg_ASONAM2014.txt`(last accessed on Oct 19, 2015)

[4] Density refers to the ratio of the number of links present in the graph to the total number of possible links.



(a) Complex Contagion Model (b) Threshold Models (c) Independent Cascade Model

Fig. 1 Agreement of simulation and theory for the three models for Digg1k dataset.

Table 2 Parameters used in the experimental validation on synthetic graphs

Methods	Parameter Sets
CCM	$\{p = 0.05, r_t = \exp(0.002n_i^{t-1} - 6)\}$
	$\{p = 0.20, r_t = \exp(0.002n_i^{t-1} - 6)\}$
	$\{p = 0.80, r_t = \exp(0.002n_i^{t-1} - 6)\}$
	$\{p = 0.05, r_t = \exp(0.0002n_i^{t-1} - 6)\}$
	$\{p = 0.20, r_t = \exp(0.0002n_i^{t-1} - 6)\}$
	$\{p = 0.80, r_t = \exp(0.0002n_i^{t-1} - 6)\}$
ICM	$p = 0.025$
	$p = 0.050$
	$p = 0.100$
	$p = 0.200$
	$p = 0.400$
	$p = 0.800$
LFM	$\{\alpha = 0.05, \beta = -2\}$
	$\{\alpha = 0.20, \beta = -2\}$
	$\{\alpha = 0.80, \beta = -2\}$

in prediction of total number of infections at time t:

$$e_t^f = \frac{|\sigma_t^* - \sigma_t|}{\sigma_t}, \tag{15}$$

where $\sigma_t = \sum_u B_{u,t}$ is obtained by simulations and $\sigma_t^* = \sum_u B_{u,t}^*$ is the predicted value obtained by Theorem 1.

Figures 2(a), 2(b) and 2(c) show how RMSE averaged over graph sizes varies with time. For ICM, the error stabilizes quickly, while for CCM and LFM the error decreases with time. The decrease is more prominent for LFM, where the error is insensitive to the parameters. Figures 3(a), 3(b) and 3(c) report average RMSE over time as a function of graph size. In all three cases, the error is very small.

Figures 4 and 5 show the variation of fractional error e_t^f with graph size and time accordingly. We note that the trend is similar to that observed for RMSE. In fact, it can be shown that the fractional error e_t^f is bounded by RMSE e_t^{rms} according to the following formula:

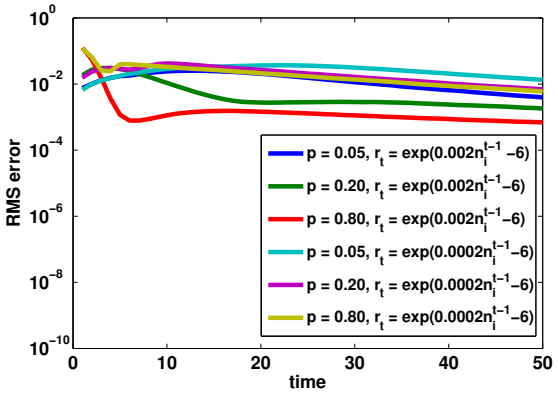
$$\frac{\sqrt{|V|}}{\sigma_t} e_t^{rms} \leq e_t^f \leq \frac{|V|}{\sigma_t} e_t^{rms}. \tag{16}$$

Figure 6 shows how RMSE varies with density. For CCM the error decreases with increasing density, whereas the error increases till some density and then falls rapidly for ICM. No clear trend is prominent in LFM; nonetheless the error is contained in a very small window (~ 0.0291 to 0.0298). RMSE curves with respect to time for different densities are shown in Figure 7. For brevity, we report results only for one parameter set for each model in this case, as we found other parameter sets produce similar trends. In all cases, the error decreases with time. The decrease becomes more apparent in high density graphs for CCM. Small density graphs require higher values of t (not shown in the figure) to reveal a similar pattern decreasing error. For ICM, the error decreases initially, but then stabilizes around a small constant value. Finally, RMSE rises initially in the case of LFM, but then falls exponentially, and is very less sensitive to the density.

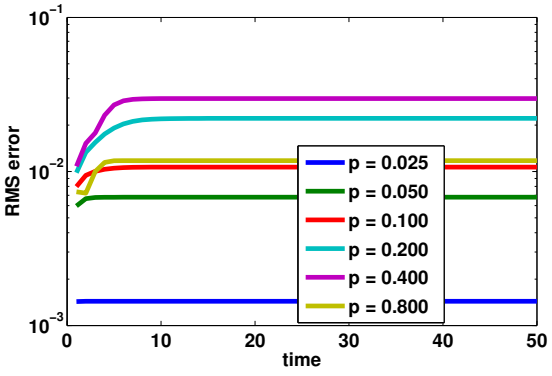
Overall, these experiments demonstrate the robustness of our model. To summarize, we find that the error remains small for different graph sizes and densities. The error is also unaffected by the various models' parameter values. This fact empirically verifies our claim that under the Unified Model of Influence, Equation 5 is a good approximation to various models with minimal computational requirements.

6 Seed Set Selection for Influence Maximization

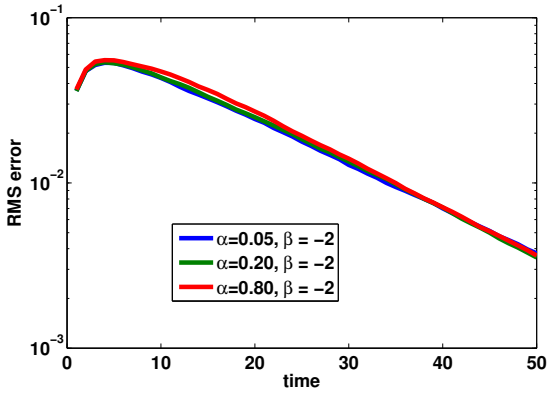
One of the fundamental problems in the diffusion literature is the influence maximization problem (Kempe et al. 2003). Let $G(V,E,W)$ be a weighted graph with vertices in V modeling the individuals and E modeling relationships between them with certain weights W representing the influence of one individual over another. The influence maximization problem is defined as follows: Given a directed



(a) CCM

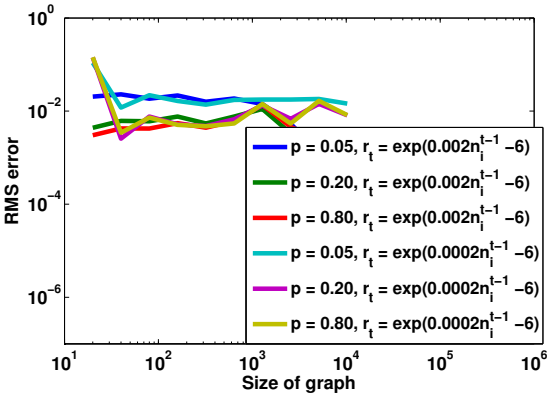


(b) ICM

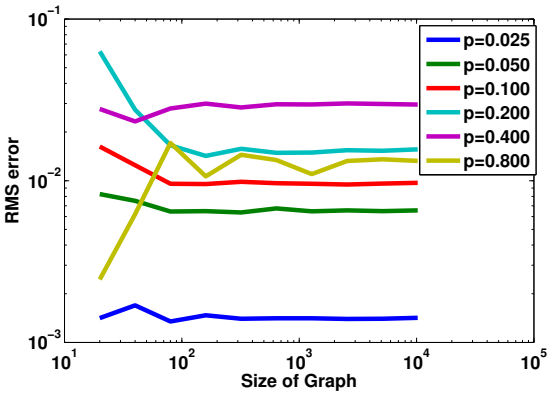


(c) LFM

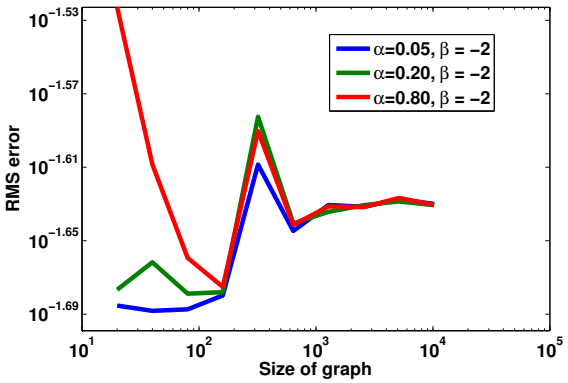
Fig. 2 RMSE as a function of time steps on synthetic graphs. RMSE is averaged over graph sizes.



(a) CCM



(b) ICM

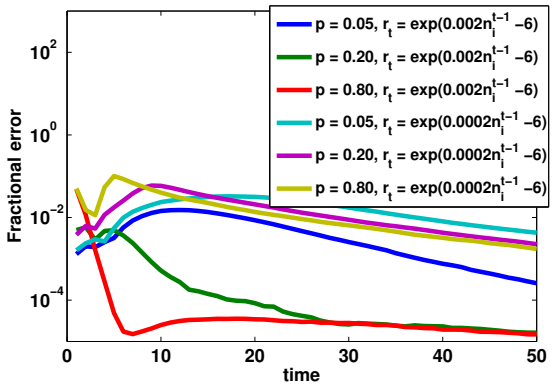


(c) LFM

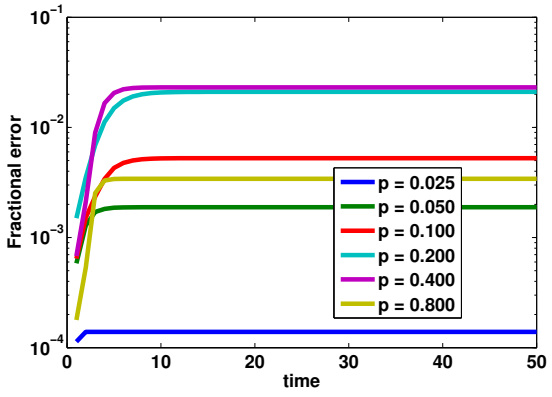
Fig. 3 RMSE averaged over time, as a function of graph size on synthetic graphs.

graph $G(V,E,W)$, a propagation model M and a positive integer $k \leq |V|$, find a seed-set $S \subseteq V$, $|S| = k$ to initiate the influence propagation, such that the expected number of influenced nodes at steady state is maximized. We claim that it is more advantageous to know how infection of a node affects the number of infections in the immediate future or in a given time frame rather than in infinite time. Therefore, we state a generalization of the problem:

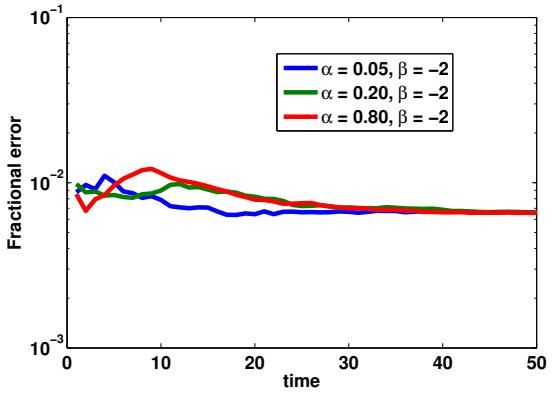
Prob. Definition 1 (Generalized Influence Maximization) *Given graph $G(V,E,W)$, infection model M, time t and positive integer k, select $S \subset V \times \mathbb{N}$ with $|S| = k$ such that $\sigma_t^M(S)$ is maximum. Here, $\sigma_t^M(S)$ denotes the expected number of infections achieved by the model M at time t with seed set S, i.e., $\forall(u,\tau) \in S$, node u is selected to initiate the infection propagation at time τ.*



(a) CCM

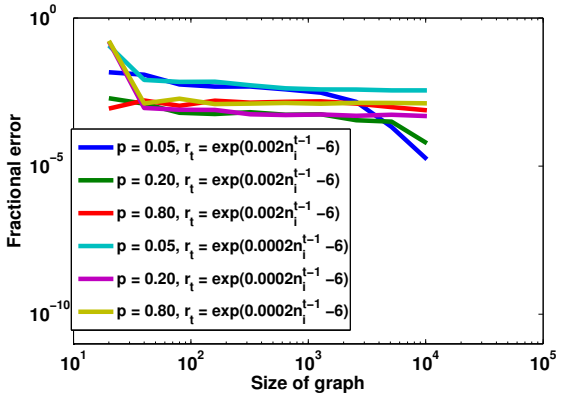


(b) ICM

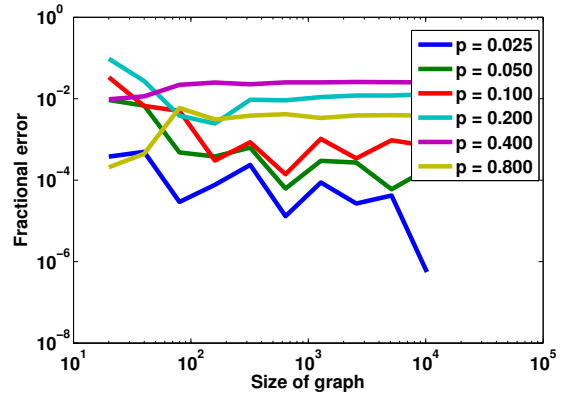


(c) LFM

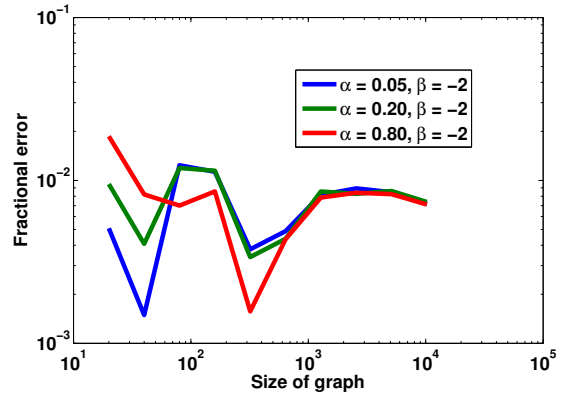
Fig. 4 Fractional error over time on synthetic graphs. Fractional error is averaged over graphs sizes.



(a) CCM



(b) ICM



(c) LFM

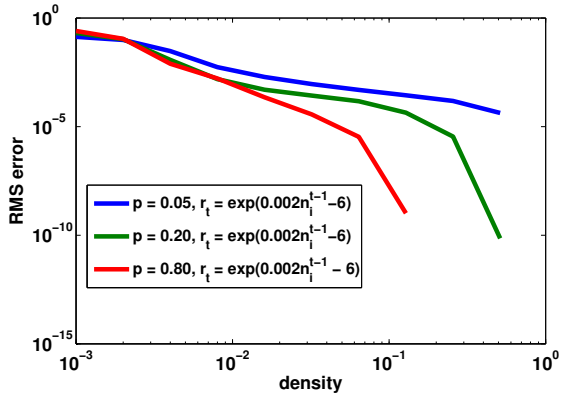
Fig. 5 Fractional error averaged over time, as a function of graph size on synthetic graphs.

6.1 Online Seed-set Selection using Unified Model

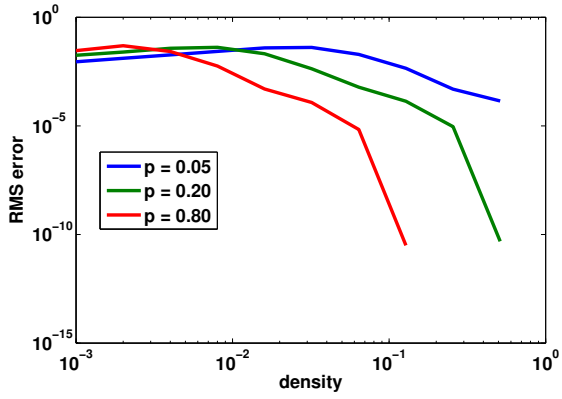
We propose Online Seed-set Selection using Unified Model (OSSUM), a greedy method based on the formula for the Unified Model (Equation 5). At each time step $t \leq k$ we select a node, infecting which would produce the maximum increase in total infection at the next time step. Algorithm 1 details the selection process. As shown in Section 4, for ICM and GLT, the infection process can be captured by the collective influence, and the Equation 5 becomes

$$B_{u,t} = 1 - (1 - r_u(t))(1 - B_{u,t-1}). \tag{17}$$

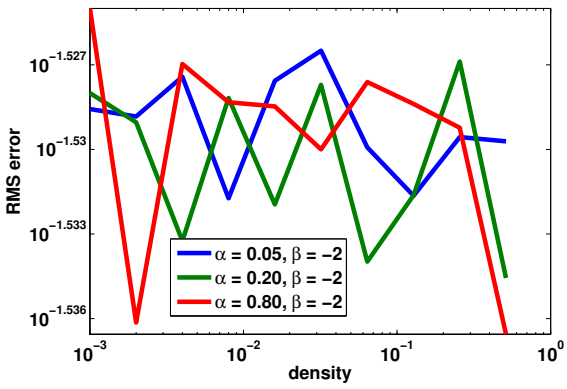
Let $r^l_u(t)$ denote the collective influence on node u if node l was manually infected at time $t-1$. Suppose $B^l_{u,t}$ denote the resultant infection probability of node u after the manual



(a) CCM

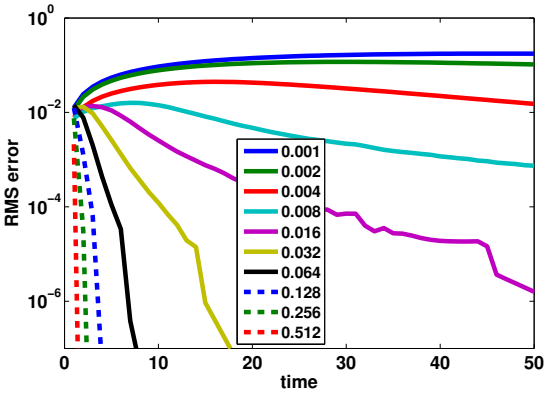


(b) ICM

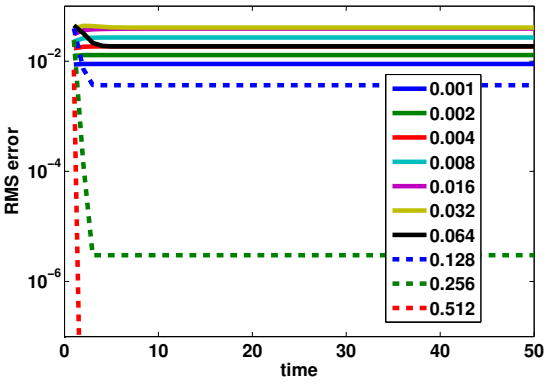


(c) LFM

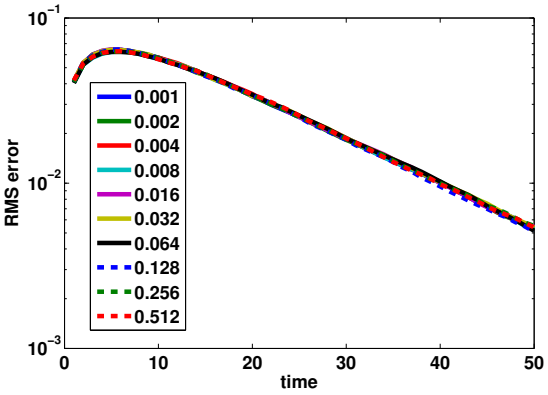
Fig. 6 RMSE on synthetic graphs for varying density, averaged over time.



(a) CCM with ($p = 0.05$, $r_t = \exp(0.002n_i^{t-1} - 6)$)



(b) ICM with $p = 0.05$



(c) LFM with $\alpha = 0.05$, $\beta = -2$

Fig. 7 RMSE over time on the synthetic graphs of size 1,000, for varying density values.

infection of node l. Then,

$$B^l_{u,t} = \begin{cases} 1-(1-r^l_u(t))(1-B_{u,t-1}) & \text{if } u \neq l \\ 1 & \text{if } u = l \end{cases} \tag{18}$$

Now, note that for both ICM and GLT, the objective function (line 4 in Algorithm 1)

$$F(l,t-1) = \sigma^M_t(S \cup (l,t-1)) - \sigma^M_t(S) = \sum_u B^l_{u,t} - \sum_u B_{u,t}$$
$$= (1-B_{l,t-1}) + \sum_{u \neq l} (r^l_u(t) - r_u(t))(1-B_{u,t-1}), \tag{19}$$

where the values of $B_{u,t}$ and $B^l_{u,t}$ are substituted from Equations 17 and 18. Using Equations 11 and 19, and after some

Algorithm 1 Online Seed-set Selection using Unified Model

```
function OSSUM(G, M, k)
    S ← ∅
    for t = 1 → k do
        arg max_l σ_t^M(S ∪ (l, t−1)) − σ_t^M(S)      ▷ Computed using Equation 5
        S ← S ∪ l
    end for
    return S
end function
```

algebraic manipulation, the objective function becomes

$$F(l,t-1) = (1-B_{l,t-1})\left(1+\sum_{u\in N_o(l)}\frac{p(1-r_u(t))(1-B_{u,t-1})}{1-pA_{l,t-1}}\right). \tag{20}$$

Computation of $F(l,t-1)$ requires $O(1+deg_o(u))$ operations, where $deg_o(u)$ is the outdegree of u. To find the node l that maximizes $F(l,t-1)$, one has to find $F(l,t-1)\forall l\in V$, which requires $O(\sum_l(1+deg_o(l))) = O(|V|+|E|)$ operations. Hence, the time complexity of finding k nodes for seed-set S for ICM using OSSUM is $O(k(|V|+|E|))$.

Similarly for LFM, the objective function can be shown to be

$$F(l,t-1) = (1-B_{l,t-1}) + \sum_{u\in N_o(l)}(1-B_{u,t-1}) \left(s\left(\alpha\sum_{j\in N_i(u)} b_{j,u}B_{j,t-1} + b_{l,u}(1-B_{l,t-1}) + \beta\right) - s\left(\alpha\sum_{j\in N_i(u)} b_{j,u}B_{j,t-1}+\beta\right)\right), \tag{21}$$

where, $s(x) = \exp(x)/(1+\exp(x)) = 1/(1+\exp(-x))$. For LFM, finding $F(l,t-1)$ requires $O(1+\sum_{u\in N_o(l)} deg_i(u))$ operations, where $deg_i(u)$ denotes the in-degree of node u. Finding $F(l,t-1)\forall l\in V$ requires $O(\sum_l(1+\sum_{u\in N_o(l)} deg_i(u)))$. Since $deg_i(u)$ counted in this expression as many times u becomes an outgoing neighbor of a node, this expression can be rewritten as

$$\sum_l(1+\sum_{u\in N_o(l)} deg_i(u)) = \sum_l(1+\sum_u deg_i(u)^2), \tag{22}$$

which is bounded by $O(|V|+|E|^2/|V|)$ (de Caen. 1998). Therefore, the time complexity of finding k nodes for seed-set S for LFM using OSSUM is $O(k(|V|+|E|^2/|V|))$.

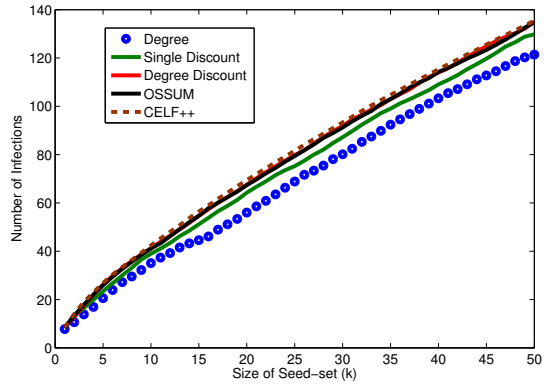
Note that the Influence Maximization problem in (Kempe et al. 2003) is a special case of Generalized Influence Maximization where $t\to\infty$ and $S\in V\times\{0\}$, i.e., all the nodes in the seed-set S are infected at $t=0$. We propose to use OSSUM which generates $S=\{(u_1,0),(u_2,1),\ldots,(u_k,k-1)\}$, and use the set $\{u_1,u_2,\ldots,u_k\}$ disregarding the time dimension. The incremental approach of OSSUM allows us to calculate the immediate effect of the chosen vertex, so that the next vertex that is included in the seed-set has least overlap with the previous one with respect to the nodes that they influence.

6.2 Experiments with Seed-set selection

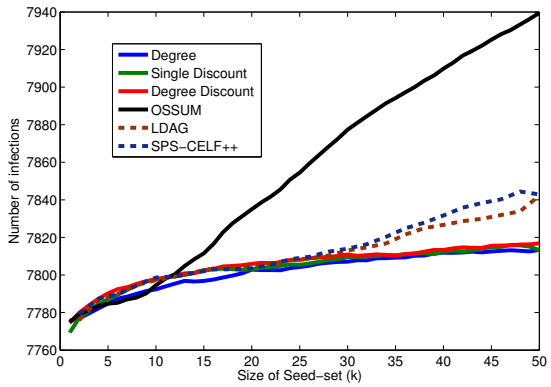
We evaluate OSSUM's ability to identify good seed-sets in a real world network. We are interested in studying its behavior in practise and compare its performance against state-of-the-art methods for Influence Maximization. We compared OSSUM against the following widely used methods for seed-set selection as baselines:

- Degree: The nodes with top k highest degrees are selected.
- Single Discount: First the highest degree node is selected and is removed from the graph. Next, the highest degree node from the remaining graph is selected. The process is continued until k nodes have been selected.
- Degree Discount (Chen et al. 2009): A heuristic designed for ICM, which performs a form of weighted discount based on the parameter p.
- CELF++ (Goyal et al. 2011a): A further optimization on CELF that exploits sub-modularity of the spreading process. Here, we use it as a baseline for ICM.
- LDAG (Chen et al. 2010b): A scalable algorithm specifically designed for Linear Threshold Model. It utilizes local DAG structures to estimate influence spread.
- SPS-CELF++ (Goyal et al. 2011b): Algorithm for influence maximization on Linear Threshold Model which performs several optimization including CELF++.

We used HEPT (Chen et al. 2009), a real world co-authorship network, where nodes are authors and a link between two nodes represents a paper co-authored by the two authors. The network consists of 15,233 nodes and 58,891 undirected edge. The baselines and our algorithm were applied on the HEPT graph for ICM and LFM. Figure 8 shows the results of these experiments. For ICM (Figure 8(a)), we set $p=0.01$ and plot σ_∞, i.e., until number of infections reach a steady state for a given seed set. OSSUM performs almost same as Degree Discount. This could be attributed to the fact that Degree Discount heuristic for ICM is based on similar assumption as our solution for the Unified Model, i.e., it assumes that every node and it's neighborhood forms a star-like network, which is equivalent to the statement that infection state of two neighbors are independent. Performance of CELF++ is also similar to that of Degree Discount, which is in agreement with the results demonstrated in (Chen et al. 2009).



(a) ICM at steady state



(b) LFM at $t = 60$

Fig. 8 Seed-set selection for Influence maximization.

A considerable deviation between OSSUM and the baselines appears in the case of LFM (Figure 8(b)). In LFM, the probability of infection of a node always remains bounded below by $s(\beta)$ (the infection probability of u at time t is $s(\alpha|In(u,t)| + \beta)$), and so at steady state ($t \to \infty$), every node becomes infected. Therefore, to study the difference between the different seed-set selection methods, we plot the number of infections achieved until $t = 60$, i.e., (σ_{60}). It can be observed that OSSUM significantly outperforms the baselines. The difference primarily arises from the fact that the infection probability of a node has a non-linear dependency on its neighborhood, which is difficult to capture by the baselines. LDAG and SPS-CELF++ perform better than the heuristic but are clearly outperformed by OSSUM. This is due to the fact these methods are specifically designed for Linear Threshold. However, OSSUM allows us to compute the infection probabilities as we select the nodes for manual infection providing a better selection algorithm, as long as the model can be fitted into the Unified Model.

Our experiment has two prominent outcomes. First, better sees-set selection can be achieved by considering considering the dynamics of influence in a network rather than solely relying on structural properties. Second, OSSUM performs as good as or better than algorithms tailored to specific influence models.

7 Conclusion

Influence analytics and diffusion prediction in online social networks have been important for many domains from marketing to public health. With the tremendous increase in the volume of data, network sizes reach millions of nodes, restricting the applicability of existing agent-based models or algorithmic solutions for diffusion prediction. In this work, we have proposed a novel, general analytical framework for influence calculation in social networks, which does not require expensive Monte-Carlo simulations. In this framework, each node has its own function of collective influence and peer influence. Both functions can vary with time, thus making our framework directly applicable to a plethora of real-world scenarios. We have shown how various popular models of diffusion constitute special cases of our model, suggesting that our formula is applicable for approximating the expected outcome of these models. Particularly, we have shown that our formula can substitute expensive simulation runs to calculate the expected probabilities of infection. We have further demonstrated that significant computation gains can be achieved using our formula instead of such models. We have validated our results using real-world social networks and a number of Erdős-Rényi random graphs.

We applied our analytical model for influence to the task of influence maximization in social networks. The process of influence, which manifests itself in a variety of domains including online social networks and social media, has been the subject of many studies. The problem of identifying a set of target nodes whose influence will maximize the overall cascade in the social network in particular has been well studied, particularly in the context of social-opinion dynamics, campaign-design and gateway finding among others. We have shown that our unified model is beneficial to seed set selection. Specifically, we have proposed OSSUM a greedy solution to the problem of Influence Maximization based on our Unified Model. We have empirically demonstrated its superiority against state of the art approaches under different influence models. Two important benefits of our approach are that a) it enables exploration and evaluation of different scenarios for large graphs under different influence models; b)considers the timing budget of influence propagation, adding a time-critical condition to the influence models. Relevant applications include but are not limited to diffusion of information in social networks, the propagation of viruses in computer networks, and the spread of epidemics in human populations.

Acknowledgements This work is supported by Chevron U.S.A. Inc. under the joint project, Center for Interactive Smart Oilfield Technologies (CiSoft), at the University of Southern California.

References

Abrahamson E, Rosenkopf L (1997) Social network effects on the extent of innovation diffusion: A computer simulation. Organization Science 8(3):289–309

Anagnostopoulos A, Kumar R, Mahdian M (2008) Influence and correlation in social networks. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, pp 7–15

Bakshy E, Rosenn I, Marlow C, Adamic L (2012) The role of social networks in information diffusion. In: Proceedings of the 21st international conference on World Wide Web, ACM, New York, NY, USA, pp 519–528

Bass FM (2004) A new product growth for model consumer durables. Manage Sci 50(12 Supplement):1825–1832

Bóta A, Krész M, Pluhár A (2013) Approximations of the generalized cascade model. Acta Cybernetica 21(1):37–51

Budak C, Agrawal D, El Abbadi A (2012) Diffusion of information in social networks: Is it all local? In: 2012 IEEE 12th International Conference on Data Mining (ICDM), pp 121–130, DOI 10.1109/ICDM.2012.74

de Caen D (1998) An upper bound on the sum of squares of degrees in a graph. Discrete Mathematics 185(1):245–248

Chelmis C, Prasanna VK (2013) The role of organization hierarchy in technology adoption at the workplace. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM, New York, NY, USA, ASONAM '13, pp 8–15

Chelmis C, Srivastava A, Prasanna VK (2014) Computational models of technology adoption at the workplace. Social Network Analysis and Mining 4(1):1–18

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

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

Chen W, Yuan Y, Zhang L (2010b) Scalable influence maximization in social networks under the linear threshold model. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on, IEEE, pp 88–97

Choi H, Kim SH, Lee J (2010) Role of network structure and network effects in diffusion of innovations. Industrial Marketing Management 39(1):170 – 177

Du N, Song L, Gomez-Rodriguez M, Zha H (2013) Scalable influence estimation in continuous-time diffusion networks. In: Advances in Neural Information Processing Systems, pp 3147–3155

Erdős P, Rényi A (1960) On the evolution of random graphs. Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5:17–61

Fan L, Wu W, Zhai X, Xing K, Lee W, Du DZ (2014) Maximizing rumor containment in social networks with constrained time. Social Network Analysis and Mining 4(1):1–10

Gionis A, Terzi E, Tsaparas P (2013) Opinion maximization in social networks. In: SDM, SIAM, pp 387–395

Gomez Rodriguez M, Leskovec J, Krause A (2010) Inferring networks of diffusion and influence. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '10, pp 1019–1028, DOI 10.1145/1835804.1835933

Goyal A, Bonchi F, Lakshmanan LV (2010) Learning influence probabilities in social networks. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, ACM, New York, NY, USA, WSDM '10, pp 241–250, DOI 10.1145/1718487.1718518

Goyal A, Lu W, Lakshmanan LV (2011a) Celf++: optimizing the greedy algorithm for influence maximization in social networks. In: Proceedings of the 20th international conference companion on World wide web, ACM, pp 47–48

Goyal A, Lu W, Lakshmanan LV (2011b) Simpath: An efficient algorithm for influence maximization under the linear threshold model. In: Data Mining (ICDM), 2011 IEEE 11th International Conference on, IEEE, pp 211–220

Granovetter M (1978) Threshold models of collective behavior. American Journal of Sociology 83(6):1420–1443

Hajibagheri A, Hamzeh A, Sukthankar G (2013) Modeling information diffusion and community membership using stochastic optimization. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM, New York, NY, USA, ASONAM '13, pp 175–182, DOI 10.1145/2492517.2492545

Hethcote HW (2000) The mathematics of infectious diseases. SIAM Review 42(4):599–653

Jacquez JA, Simon CP (1993) The stochastic si model with recruitment and deaths i. comparison with the closed sis model. Mathematical Biosciences 117(1-2):77–125

Kamp C (2010) Untangling the interplay between epidemic spread and transmission network dynamics. PLoS Computational Biology 6(11):e1000,984

Kelman HC (1961) Processes of opinion change. Public opinion quarterly 25(1):57–78

Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, pp 137–146

Kleinberg J (2007) Cascading behavior in networks: Algorithmic and economic issues. In: Nisan N, Roughgarden T, Tardos E, Vazirani VV (eds) Algorithmic Game Theory, Cambridge University Press

Lahiri M, Cebrian M (2010) The genetic algorithm as a general diffusion model for social networks. In: AAAI

Leskovec J, Adamic LA, Huberman BA (2006) The dynamics of viral marketing. In: Proceedings of the 7th ACM conference on Electronic commerce, ACM, New York, NY, USA, pp 228–237

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

Lin YR, Sun J, Castro P, Konuru R, Sundaram H, Kelliher A (2009) Metafac: Community discovery via relational hypergraph factorization. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '09, pp 527–536, DOI 10.1145/1557019.1557080

Lu Z, Zhang W, Wu W, Kim J, Fu B (2012) The complexity of influence maximization problem in the deterministic linear threshold model. Journal of combinatorial optimization 24(3):374–378

Myers SA, Zhu C, Leskovec J (2012) Information diffusion and external influence in networks. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '12, pp 33–41, DOI 10.1145/2339530.2339540

Newman ME (2002) Spread of epidemic disease on networks. Physical review E 66(1):016,128

Shakarian P, Eyre S, Paulo D (2013) A scalable heuristic for viral marketing under the tipping model. Social Network Analysis and Mining 3(4):1225–1248

Srivastava A, Chelmis C, Prasanna VK (2014) Influence in social networks: A unified model? In: Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (To Appear), ASONAM '14

Srivastava A, Chelmis C, Prasanna VK (2015 (Accepted)) Computational models for cascades in massive graphs: How to spread a rumor in parallel. In: Bader DA (ed) Parallel Graph Algorithms, Chapman and Hall/CRC Computational Science

Subbian K, Sharma D, Wen Z, Srivastava J (2013) Finding influencers in networks using social capital. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM, New York, NY, USA, ASONAM '13, pp 592–599, DOI 10.1145/2492517.2492552

Tang J, Sun J, Wang C, Yang Z (2009) Social influence analysis in large-scale networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '09, pp 807–816, DOI 10.1145/1557019.1557108

Tong H, Papadimitriou S, Faloutsos C, Philip SY, Eliassi-Rad T (2010) Basset: Scalable gateway finder in large graphs. In: Advances in Knowledge Discovery and Data Mining, Springer, pp 449–463

Valente TW (1996) Social network thresholds in the diffusion of innovations. Social Networks 18(1):69–89

Xu W, Lu Z, Wu W, Chen Z (2014) A novel approach to online social influence maximization. Social Network Analysis and Mining 4(1):1–13

Yang J, Leskovec J (2010) Modeling information diffusion in implicit networks. In: Proceedings of the 2010 IEEE International Conference on Data Mining, IEEE Computer Society, Washington, DC, USA, pp 599–608