Influence Maximization in Continuous Time Diffusion Networks

Manuel Gomez-Rodriguez[1,2]
Bernhard Schölkopf[1]

[1]MPI for Intelligent Systems and [2]Stanford University

MANUELGR@STANFORD.EDU
BS@TUEBINGEN.MPG.DE

Abstract

The problem of finding the optimal set of source nodes in a diffusion network that maximizes the spread of information, influence, and diseases in a limited amount of time depends dramatically on the underlying temporal dynamics of the network. However, this still remains largely unexplored to date. To this end, given a network and its temporal dynamics, we first describe how continuous time Markov chains allow us to analytically compute the average total number of nodes reached by a diffusion process starting in a set of source nodes. We then show that selecting the set of most influential source nodes in the continuous time influence maximization problem is NP-hard and develop an efficient *approximation algorithm* with provable near-optimal performance. Experiments on synthetic and real diffusion networks show that our algorithm outperforms other state of the art algorithms by at least ∼20% and is robust across different network topologies.

1. Introduction

In recent years, there has been an increasing effort in uncovering, understanding, and controlling diffusion and propagation processes in a broad range of domains: information propagation (Leskovec et al., 2007), social networks (Kempe et al., 2003), viral marketing (Richardson & Domingos, 2002), and epidemiology (Wallinga & Teunis, 2004). Diffusion networks have raised many research problems, ranging from network inference (Gomez-Rodriguez et al., 2010; 2011) to influence spread maximization (Kempe et al., 2003). In this article, we pay attention to the latter problem, and we propose a method for continuous time influence maximization that accounts for the temporal dynamics of diffusion networks.

Influence spread maximization tackles the problem of selecting the most influential source node set of a given size in a diffusion network. A diffusion process that starts in such an influential set of nodes is expected to reach the greatest number of nodes in the network. In information propagation, the problem reduces to choosing the set of blogs and news media sites that together are expected to spread a piece of news to the greatest number of sites. In viral marketing, it consists of identifying the most influential set of *trendsetters* that together may influence the greatest number of customers. Finally, in epidemiology, the influence maximization problem reduces to finding the set of individuals that together are most likely to spread an illness or virus to the greatest percentage of the population. In this latter case, the solution of the influence maximization problem helps towards developing vaccination and quarantine policies.

In our work, we build on the fully continuous time model of diffusion recently introduced by Gomez-Rodriguez et al. (2011). This model accounts for temporally heterogeneous interactions within a diffusion network – it allows information (or influence) to spread at different rates across different edges, as shown in real-world examples. We first describe how, given a set of source nodes, we can compute the average total number of infected nodes analytically using the work of Kulkarni (1986). The key observation is that the infection time of a node in a network with stochastic edge lengths is the length of the stochastic shortest path from the source nodes to the node. Later, we show that finding the optimal influential set of source nodes in the continuous time influence maximization problem is a NP-hard problem. We then provide an *approximation algorithm* that finds a suboptimal set of source nodes with *provable guarantees* in terms of the average total number of infected nodes.

Related work. Richardson & Domingos (2002) were the first to study influence maximization as an algorithmic problem, motivated by marketing applications. In their work, they proposed heuristics for choosing a set of influential customers with a large overall effect on a network, and methods to infer the influence of each customer were de-

veloped. Kempe et al. (2003) posed influence maximization in a social network as a discrete optimization problem. They showed that the optimal solution is NP-hard for several models of influence, and obtained the first provable approximation guarantees for efficient algorithms based on a natural diminishing property of the problem, submodularity. Since then there have been substantial developments that build on their seminal work. Efficient influence maximization that uses heuristics to speed up the optimization problem has been proposed (Chen et al., 2009; 2010) and influence maximization has been studied on the context of competing cascades (Bharathi et al., 2007) or under additional constraints (Goyal et al., 2010).

However, to the best of our knowledge, previous work on influence maximization has ignored the underlying temporal dynamics governing diffusion networks – once a transmission occurs, it always occurs at the same rate or temporal scale. In contrast, we consider heterogeneous pairwise transmission rates, found in many real-world examples. In information propagation, news media sites and professional bloggers typically report news faster than people that maintain personal blogs. In epidemiology, people meet each other with different frequencies and then the pairwise transmission rates between individuals within a population differ. Finally, in viral marketing, some customers make up their minds about a product or service quicker than others, and then pass recommendations on at different rates.

The main contribution of our work is twofold. First, it considers a novel continuous time formulation of the influence maximization problem in which information or influence can spread at different rates across different edges, as in real-world examples. Second, this continuous time approach allows us to analytically compute and efficiently optimize the influence (*i.e.*, average total number of infections), avoiding the use of heuristics (Chen et al., 2010; 2009) or Monte Carlo simulations (Kempe et al., 2003).

2. Problem formulation

In this section, we build on the fully continuous time model of diffusion recently proposed by Gomez-Rodriguez et al. (2011). We start by describing how the diffusion model accounts for pairwise interactions and then continue discussing some basic assumptions about diffusion processes. We conclude with a statement of the continuous time influence maximization problem.

Pairwise transmission likelihood. In a diffusion network, we first need to model the pairwise interactions between nodes. We pay attention to the general case in which different pairwise interactions between nodes in the network occur at different rates. Define $f(t_j|t_i; \alpha_{i,j})$ as the conditional likelihood of transmission between a node i and

a node j, where t_i and t_j are infection times and $\alpha_{i,j}$ is the transmission rate. We assume that the likelihood depends on the pairwise transmission rate $\alpha_{i,j}$ and the time difference $(t_j - t_i)$ (*i.e.*, it is time shift invariant). Moreover, a node cannot be infected by a node infected later in time (*i.e.*, $t_j > t_i$) and as $\alpha_{i,j} \to 0$, the expected transmission time becomes arbitrarily long.

In the remainder of the paper, we consider the exponential distribution $f(t_j|t_i; \alpha_{i,j}) \propto e^{-\alpha_{i,j}(t_j - t_i)}$ to model pairwise interactions for the sake of simplicity. The exponential model is a well-known parametric model for modeling diffusion and influence in social and information networks (Gomez-Rodriguez et al., 2010). However, our results can easily be extended to diffusion networks with phase-type pairwise transmission likelihoods. This is important since the set of phase-type distributions is dense in the field of all positive-valued distributions and it can used to approximate power-laws, which have been also used for modeling diffusions in social networks (Gomez-Rodriguez & Schölkopf, 2012), Rayleigh distributions, which have been used in epidemiology (Wallinga & Teunis, 2004), and also subprobability distributions, which enable us to describe two step traditional diffusion models (Kempe et al., 2003), in which with probability $(1 - \beta)$ an infection may never occur.

Continuous time diffusion process. We consider diffusion and propagation processes that occur over static networks with known (or inferred) connectivity and transmission rates. A diffusion process starts when a source node set A becomes infected at time $t = 0$ by action of an external source to the network. Then, source nodes try to infect their children (*i.e.*, neighbors that they can reach directly through an outgoing edge). Once a child i gets infected at time t_i, it tries to infect her own children, and so on. For some pairwise transmission likelihoods, it may happen that $t_i \to \infty$ and child i is never infected. Here, we assume that a node i becomes infected as soon as one of its parents (*i.e.*, neighbors that are able to reach node i through an outgoing edge) infects it, and later infections by other parents do not contribute anymore towards the evolution of the diffusion process. As a consequence of this assumption, at any time $t \geq 0$ there may be some nodes and edges in the network that are useless for the spread of the information (be it in the form of a meme, a sales decision or a virus) towards a specific node n. If these nodes get infected and transmit the information to other nodes, this information can only reach n through previously infected nodes. Therefore, the infection time t_n of node n does not depend on these nodes.

Finally, given a diffusion process that started in the set of source nodes A, we define $N(A; T)$ as the number of nodes infected up to time T and then define the influence function
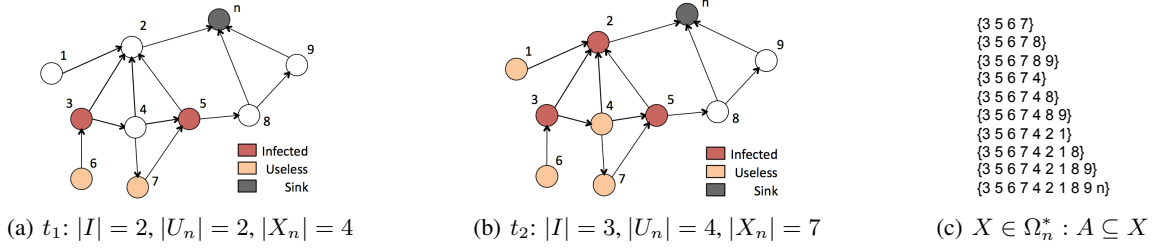
(a) t_1: $|I| = 2, |U_n| = 2, |X_n| = 4$

(b) t_2: $|I| = 3, |U_n| = 4, |X_n| = 7$

(c) $X \in \Omega_n^* : A \subseteq X$

Figure 1. Panels (a,b): Sets of infected nodes (I; in red) and useless nodes (U_n; in orange) at two different times for a diffusion process that starts in the source node set $A = \{3, 5\}$ relative to a particular sink node (n; in black) . Any path from a useless node to the sink node is *blocked* by an infected node. The set of disabled (X_n) nodes is simply the union of the sets of infected and useless nodes. Panel (c): Sets of disabled nodes $X \in \Omega_n^*$ such that $A \subseteq X$. They represent the states that we need to describe the temporal evolution of a diffusion process towards the sink node n that starts in the set of sources A.

$\sigma(A; T)$ as the average total number of nodes infected up to time T, *i.e.*, $\sigma(A; T) = \mathbb{E}N(A; T)$.

Continuous time influence maximization problem. Our goal is to find the set of source nodes A in a diffusion network G that maximizes the influence function $\sigma(A; T)$. In other words, the set of source nodes A such that a diffusion process in G reaches, on average, the greatest number of nodes before a window cut off T. Thus, we aim to solve:

$$A^* = \underset{|A| \leq k}{\operatorname{argmax}} \, \sigma(A; T), \qquad (1)$$

where the source set A is the variable to optimize and the time horizon T and the source set cardinality k are constants.

3. Proposed algorithm

We start this section by describing how to evaluate the influence function $\sigma(A; T)$ for any set of sources A in a network G using the work of Kulkarni (1986). The key observation is that the infection time of a node in a network with stochastic edge lengths is the length of the stochastic shortest path from the source nodes to the node. Then, we show that the continuous time influence maximization problem defined by Eq. 1 is NP-hard. Finally, we show how to efficiently find a *provable near-optimal* solution to our maximization problem by exploiting a natural diminishing returns property of our objective function.

Evaluating the influence. The influence function depends on the probability of infection of every node in the network as follows:

$$\sigma(A; T) = \mathbb{E}N(A; T) = \sum_{n=1}^{N} P(t_n \leq T | A), \qquad (2)$$

where t_n is the infection time of node n, A is the set of source nodes, and T is the time horizon or time window

cut-off. Therefore, we need to compute the probability of infection $P(t_n \leq T | A)$ for each node n in the network. Note that whenever $n \in A$, the probability of infection $P(t_n \leq T | A)$ is trivially 1. We will refer to node n as sink node.

Revisiting the basic assumptions about a diffusion process that we presented in Section 2, we recall some definitions to describe its temporal evolution as in Kulkarni (1986). Given a diffusion network $G = (V, E)$, a set of nodes $B \subset V$, and a node $n \in V$, we define the set of nodes blocked by or dominated by B:

$$S_n(B) = \{u \in V : \text{any path from } u \text{ to } n \text{ in } G \text{ visits}$$
$$\text{at least one node in } B\}.$$

By definition, $B \subseteq S_n(B)$ and $S_n(S_n(B)) = S_n(B)$. We now define the set Ω_n^* as:

$$\Omega_n^* = \{X \subset V : X = S_n(X)\}.$$

In words, all nodes in $X \in \Omega_n^*$ block only themselves relative to the sink node n. We can find all sets in Ω_n^* efficiently (Georgiadis et al., 2006; Provan & Shier, 1996). In particular, we are able to find each $X \in \Omega_n^*$ in time $O(|V|)$. However, in dense networks, $|\Omega_n^*|$ can be exponentially large and lead to a worst-case non polynomial time algorithm. In order to illustrate this, we compute $\max_n |\Omega_n^*|$ across $1,000$ random source sets with $|S| = 5$ and $|S| = 10$ for several 256-node hierarchical networks of increasing network density. We observe that $\max_n |\Omega_n^*| < 85$ for all networks up to 2 edges per node in average. However, $\max_n |\Omega_n^*|$ grows quickly for higher network densities (*e.g.*, $\max_n |\Omega_n^*| < 7750$ for a network with 2.5 edges per node in average). In order to overcome this drawback, we will propose several speed-ups (LTP and LSN) that provide approximate solutions or sparsify the networks as in Mathioudakis et al. (2011).

Given a diffusion process that starts in a set of source nodes A, a sink node n and any time $t \geq 0$, we denote the set

Algorithm	$|\mathbf{A}|$	$\sigma(\mathbf{A}; \mathbf{0.1})$	$\sigma(\mathbf{A}; \mathbf{0.5})$	$\sigma(\mathbf{A}; \mathbf{1.0})$
Enumeration	1	3.05	8.95	13.90
	3	8.04	18.01	21.70
	5	11.60	22.17	25.59
INFLUMAX	1	3.05	8.95	13.90
	3	8.04	18.01	21.70
	5	11.60	22.17	25.59
Greedy	1	3.05	7.70	9.31
	3	6.18	12.70	15.88
	5	8.62	16.05	19.70
PMIA	1	1.20	1.67	1.89
	3	4.90	11.67	16.80
	5	8.90	18.03	22.02
SP1M	1	2.15	8.04	11.66
	3	4.88	10.75	13.14
	5	7.96	13.95	16.16
Random	1	1.61	3.77	5.34
	3	4.63	9.29	11.84
	5	7.39	13.36	16.04

Table 1. Influence $\sigma(A; T)$ that enumeration, INFLUMAX and several other baselines achieve in a small Kronecker core-periphery network with 35 nodes and 39 edges for different time horizon values T and number of sources $|A|$. INFLUMAX always achieves the optimal influence that exhaustive search gives but several order of magnitude faster.

of infected nodes as $I(t|A)$, the set of useless nodes as $U_n(t|A)$, and the set of *disabled* nodes (*i.e.*, infected or useless) as $X_n(t|A)$. Useless nodes are nodes that if they get infected and transmit the information to other nodes, this information can only reach the sink node n through previously infected nodes. Figures 1(a) and 1(b) illustrate the set of infected nodes (I) and the set of useless nodes (U_n) for a diffusion process in a small network at two different times. Note that the set of disabled nodes (X_n) is composed of the sets of infected (I) and useless nodes (U_n). By definition of $S_n(\cdot)$, $U_n(t|A) = S_n(I(t|A))\backslash I(t|A)$ and $X_n(t|A) = S_n(I(t|A))$. Now, by studying the temporal evolution of $X_n(t|A)$ we will be able to compute $P(t_n \leq T|A)$.

First, for a diffusion process that starts in the set of source nodes A, it can be shown that the set of disabled nodes $X_n(t|A)$ at any time $t \geq 0$ belongs to Ω_n^*.

Theorem 1. *(Kulkarni (1986)) Given a set of source nodes A, a sink node n and any time $t \geq 0$, $X_n(t|A) \in \Omega_n^*$.*

Figure 1(c) enumerates all sets of disabled nodes $X \in \Omega_n^*$ such that $A \subseteq X$ for the small network depicted in Figures 1(a) and 1(b). They represent the states that we need to describe the evolution of a diffusion process that starts in the set of sources A relative to the sink node n. Now, assuming independent pairwise exponential transmission likelihoods in the diffusion network, the following Th. applies:

Theorem 2. *(Kulkarni (1986)) Given a set of source nodes A, a sink node n and independent pairwise exponential*

transmission likelihoods $f(t_j|t_i; \alpha_{i,j})$, $\{X_n(t|A), t \geq 0\}$ is a continuous time Markov chain (CTMC) with state space $\{X : X \in \Omega_n^*, A \subseteq X\}$ and infinitesimal generator matrix $Q = [q(D, B)]$ ($D, B \in \{X : X \in \Omega_n^*, A \subseteq X\}$) given by:

$$q(D, B) = \begin{cases} \sum_{(i,j) \in C_v(D)} \alpha_{i,j} & \exists v : B = S_n(D \cup \{v\}), \\ -\sum_{(i,j) \in C(D)} \alpha_{i,j} & B = D, \\ 0 & otherwise. \end{cases}$$

where $C(D)$ is the unique minimal cut between D and $\bar{D} = V \backslash D$ and $C_v(D) = \{(u, v) \in C(D)\}$.

Finally, let t_n be the *length* of the fastest (shortest) directed path from any of the nodes in A to the sink node n in the directed acyclic graph (DAG) induced by the diffusion process on network G. By construction of the CTMC $\{X_n(t|A), t \geq 0\}$ in Theorem 2,

$$t_n = \min\{t \geq 0 : X_n(t|A) = S_N | X_n(0|A) = S_1\},$$

where S_1 and S_N denote respectively the first and last state of the CTMC. The *length* of the fastest (shortest) path is thus equivalent to the time until the CTMC $\{X_n(t|A), t \geq 0\}$ becomes absorbed in the final state S_N starting from state S_1 (*i.e.*, the state in which only the source nodes in A are infected). Then, computing the probability of infection of the sink node $P(t_n \leq T|A)$ reduces to computing the distribution of time of the sink state of the CTMC. Such distributions are called continuous phase-type distributions. Their generator matrix Q and the cumulative density function satisfy (Gikhman & Skorokhod, 2004):

$$P(t_n \leq T|A) = 1 - [\mathbf{10}]' e^{ST} \mathbf{1}, \text{ where } Q = \begin{bmatrix} S & S^0 \\ \mathbf{0}' & 0 \end{bmatrix},$$

where e^{ST} denotes the exponential matrix, S is the submatrix of Q that results from removing the column and row associated to the last state S_N, and $S^0 = -S\mathbf{1}$. By construction, $\{X_n(t|A), t \geq 0\}$ has the structure of a DAG and it is usually sparse. Then, S is upper triangular, sparse and e^{ST} can be computed efficiently.

As noted in Kulkarni (1986), this approach can be easily extended to diffusion networks with phase-type transmission likelihoods, which can approximate power-laws, Rayleigh or subprobability distributions.

Maximizing the influence. We have shown how to analytically evaluate our objective function $\sigma(A; T)$ for any set of sources A. However, optimizing $\sigma(A; T)$ with respect to the set of sources A seems to be a cumbersome task and naive brute-force search over all k node sets is intractable even for relatively small networks. Indeed, we cannot expect to find the optimal solution to the continuous time influence maximization problem defined by Eq. 1 since it is NP-hard:
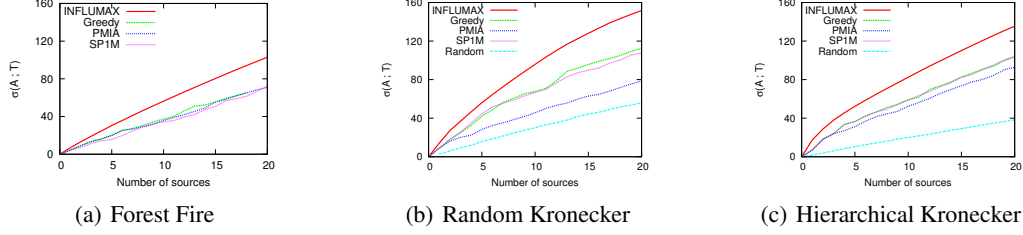
Figure 2. Panels plot influence $\sigma(A;T)$ (*i.e.*, average number of infected nodes) for $T = 1$ and transmission rates drawn from $\alpha \sim U(0,5)$ against number of sources. (a): 1,024 node Forest Fire network. (b): 512 node random Kronecker network. (c): 1,024 node hierarchical Kronecker network. The proposed algorithm INFLUMAX outperforms all other methods typically by at least 20%.

Theorem 3. *Given a network $G = (V, E)$, a set of nodes $A \subseteq V$ and a time horizon T, the continuous time influence maximization problem defined by Eq. 1 is NP-hard.*

Proof. If we let $T \rightarrow \infty$, the independent cascade model is a particular case of our continuous time diffusion model. Then, our problem is NP-hard by applying Th. 2.4 in Kempe et al. (2003). □

By construction, $\sigma(\emptyset, T) = 0$ and $\sigma(A;T) \geq 0$. It also follows trivially that $\sigma(A;T)$ is monotonically nondecreasing in the set of source nodes A, *i.e.*, $\sigma(A;T) \leq \sigma(A';T)$, whenever $A \subseteq A'$. Fortunately, we now show that the objective function $\sigma(A;T)$ is a submodular function in the set of source nodes A. A set function $F : 2^W \rightarrow \mathbb{R}$ mapping subsets of a finite set W to the real numbers is submodular if whenever $A \subseteq B \subseteq W$ and $s \in W \setminus B$, it holds that $F(A \cup \{s\}) - F(A) \geq F(B \cup \{s\}) - F(B)$, i.e., adding s to the set A provides a bigger marginal gain than adding s to the set B. By this natural diminishing returns property, we are able to find a *provable near-optimal* solution to our problem:

Theorem 4. *Given a network $G = (V, E)$, a set of nodes $A \subseteq V$ and a time horizon T, the influence function $\sigma(A;T)$ is a submodular function in the set of nodes A.*

Proof. We follow the proof of Th. 2.2 in Kempe et al. (2003). For simplicity, we assume that the infection time of all nodes in A is $t = 0$; the results generalize trivially. Consider the probability distribution of all possible time differences between each pair of nodes in the network. Thus, given a sample $\mathbf{\Delta t}$ in the probability space, we define $\sigma_{\mathbf{\Delta t}}(A;T)$ as the total number of nodes infected in a time less than or equal to T for $\mathbf{\Delta t}$.

Define $R_{\mathbf{\Delta t}}(k;T)$ as the set of nodes that can be reached from node k in a time shorter than T. It follows trivially that $\sigma_{\mathbf{\Delta t}}(A;T) = |\cup_{k \in A} R_{\mathbf{\Delta t}}(k;T)|$. Define $R_{\mathbf{\Delta t}}(k|N;T)$ as the set of nodes that can be reached from node k in a time shorter than T and at the same time cannot be reached in a time shorter than T from any node in

the set of nodes $N \subseteq V$. It follows that $|R_{\mathbf{\Delta t}}(k|N;T)| \geq |R_{\mathbf{\Delta t}}(k|N';T)|$ for the sets of nodes $N \subseteq N'$.

Consider now the sets of nodes $A \subseteq A' \subseteq V$, and a node a such that $a \notin A'$. Using the definition of submodularity,

$$\sigma_{\mathbf{\Delta t}}(A \cup \{a\};T) - \sigma_{\mathbf{\Delta t}}(A;T) = |R_{\mathbf{\Delta t}}(a|A;T)|$$
$$\geq |R_{\mathbf{\Delta t}}(a|A';T)|$$
$$= \sigma_{\mathbf{\Delta t}}(A' \cup \{a\};T) - \sigma_{\mathbf{\Delta t}}(A';T),$$

and thus $\sigma_{\mathbf{\Delta t}}(A;T)$ is submodular. Then, it follows that $\sigma(A;T)$ is also submodular. □

A well-known approximation algorithm to maximize monotonic submodular functions is the *greedy algorithm*. It adds nodes to the source node set A sequentially. In step k, it adds the node a which maximizes the *marginal gain* $\sigma(A_{k-1} \cup \{a\};T) - \sigma(A_{k-1};T)$. The greedy algorithm finds a source node set which achieves at least a constant fraction $(1 - 1/e)$ of the optimal (Nemhauser et al., 1978).

Moreover, we can also use the submodularity of $\sigma(A;T)$ to acquire a tight *online* bound on the solution quality obtained by *any* algorithm:

Theorem 5 (Leskovec et al. (2007))**.** *For a source set $\hat{A} \subseteq V$ with k sources and a node $a \in V \setminus \hat{A}$, let $\delta_a = \sigma(\hat{A} \cup \{a\};T) - \sigma(\hat{A};T)$ and $a_1, \ldots a_k$ be the sequence of k nodes with δ_a in decreasing order. Then, $\max_{|A| \leq k} \sigma(A;T) \leq \sigma(\hat{A};T) + \sum_{i=1}^{k} \delta_{a_i}$.*

Lazy evaluation (Leskovec et al., 2007) can be employed to speed-up the computation of the on-line bound for our algorithm, that we will refer as INFLUMAX.

Speeding-up INFLUMAX. We can speed up our algorithm by implementing the following speed-ups:

Lazy evaluation (LE, Leskovec et al. (2007)): it dramatically reduces the number of evaluations of marginal gains by exploiting the submodularity of $\sigma(A;T)$.

Localized source nodes (LSN): for each node n, we speed up the computation of $P(t_n \leq T|A)$ by ignoring any $a \in A$

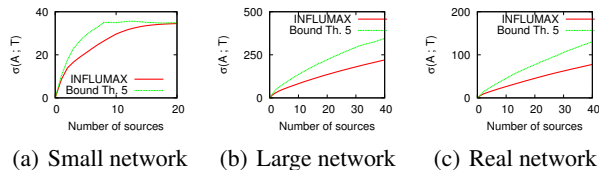(a) Small network   (b) Large network   (c) Real network

Figure 3. Influence $\sigma(A; T)$ achieved by INFLUMAX in comparison with the online upper bound from Theorem 5 for $T = 1$. (a) 35-node core-periphery Kronecker network. (b) 1,024 node hierarchical Kronecker network. (c) 1,000 node real diffusion network that we infer from hyperlinks cascades ($T = 1$).

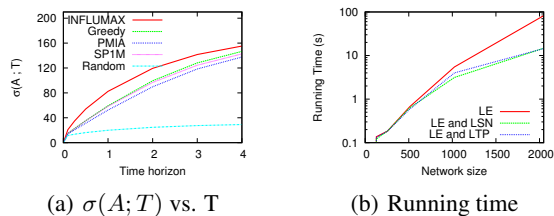(a) $\sigma(A; T)$ vs. T   (b) Running time

Figure 4. Panels show (a) influence $\sigma(A; T)$ vs. time horizon and (b) average computation time per source added for INFLUMAX implemented with (i) lazy evaluation (LE), (ii) LE and localized source nodes (LSN, $m = 6$), and (iii) LE and limited transmission paths (LTP, $m = 6$) against number of nodes.

whose shortest path to n traverses more than m nodes.

Limited transmission paths (LTP): for each node n, we speed up the computation of $P(t_n \leq T|A)$ by ignoring any path from $a \in A$ to n that traverses more than m nodes.

LSN and LTP should be used with care since they provide an approximate $P(t_n \leq T|A)$. In the remainder of this article, if not specified, we run INFLUMAX with LE but avoid using LSN and LTP.

4. Experimental evaluation

We evaluate our algorithm INFLUMAX on (i) synthetic networks that mimic the structure of real networks and on (ii) real networks inferred from the MemeTracker dataset[1] by using NETRATE's public implementation (Gomez-Rodriguez et al., 2011). We show that INFLUMAX outperforms three state of the art algorithms: the traditional greedy algorithm (Kempe et al., 2003), PMIA (Chen et al., 2010) and SP1M (Chen et al., 2009).

4.1. Experiments on synthetic data

Experimental setup. We perform experiments on two types of synthetic networks that mimic the structure of directed social networks: Kronecker (Leskovec et al., 2010) and Forest Fire (scale free) (Barabási & Albert, 1999) networks. We consider three types of Kronecker networks with very different structure: random (Erdős & Rényi, 1960) (parameter matrix [0.5, 0.5; 0.5, 0.5]), hierarchial (Clauset et al., 2008) ([0.9, 0.1; 0.1; 0.9]) and core-periphery (Leskovec et al., 2010) ([0.9, 0.5; 0.5, 0.3]).

First, we generate a network G using one of the network models cited above. Then, we draw a transmission rate for each edge $(j, i) \in G$ from a uniform distribution. We can control the transmission rate variance across edges in the network by tuning the parameters values of the distribution. In social networks, transmission rates model how fast

---
[1]Data available at `http://memetracker.org`

information spreads across the network. Given G and the transmission rates $\alpha_{j,i}$, our aim is to find the most influential subset of k nodes, *i.e.*, the subset of nodes that maximizes the spread of information up to a time T. In the traditional greedy algorithm, PMIA and SP1M, we ignore any of the transmission rates and consider all network edges to be active with probability 1, *i.e.*, we do not consider the temporal dynamics. We did not need to use Montecarlo in the traditional greedy algorithm since we assume all edges to be always active.

Solution quality. First, we compare INFLUMAX to exhaustive search and several state of the art algorithms on a small network. By studying a small network in which exhaustive search can be run, we are able estimate exactly how far INFLUMAX is from the NP-hard to find optimum. We then compare INFLUMAX to the state of the art on different large networks. Running exhaustive search on large networks is computationally too expensive and we compute instead the tight on-line bound from Th. 5.

We compare INFLUMAX to several state of the art methods on a small core-periphery Kronecker network with 35 nodes and 39 edges and transmission rates drawn from a uniform distribution $\alpha \sim U(0, 10)$. We summarize the results in Table 1. In addition to INFLUMAX and three state of the art methods, we also run a baseline that simply chooses the set of sources randomly. For all methods, we compute the influence they achieve by evaluating Eq. 2 for the set of sources selected by them. Surprisingly, INFLUMAX achieves in most cases the optimal influence that exhaustive search gives but several order of magnitude faster. In other words, the solution given by INFLUMAX may be in practice much closer to the NP-hard to find optimum than $(1 - 1/e)$, the theoretical guarantee given by Nemhauser et al. (1978), and it outperforms other methods by 20%.

Now, we focus on different large synthetic networks. Figure 2 shows the average total number of infected nodes against number of sources that INFLUMAX achieves in
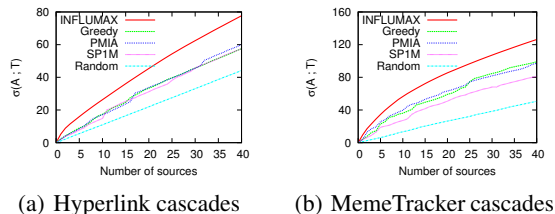
(a) Hyperlink cascades   (b) MemeTracker cascades

Figure 5. Influence $\sigma(A; T)$ for time horizon $T = 1$ against number of sources for (a) a 1,000 node real diffusion network that we infer from hyperlinks cascades and (b) a 1,000 node real diffusion network that we infer from MemeTracker cascades. The proposed algorithm INFLUMAX outperforms all other methods by 20-25%.

comparison with the other methods on a 512 node random Kronecker network, a 1,024 node hierarchical Kronecker network and a 1,024 node Forest Fire (scale free) network. All three networks have approximately 2 edges in average per node. We set the time horizon to $T = 1.0$ and the transmission rates are drawn from a uniform distribution $\alpha \sim U(0, 5)$. INFLUMAX typically outperforms other methods by at least 20% by exploiting the temporal dynamics of the network. We also compare INFLUMAX with the on-line bound from Th. 5. Fig. 3 shows the average number of infected nodes against number of sources that INFLUMAX achieves in comparison with the on-line bound for the small core-periphery Kronecker network and the large hierarchical Kronecker network that we used previously. If we pay attention to the value of the bound on the small network for source set sizes significantly smaller than the number of nodes in the network, we observe that the bound value on the influence is not as close to the optimal value given by exhaustive search as we could expect. That means that although the bound is not very tight on the large network, we may be actually achieving in practice an almost optimal value on that network too.

Influence vs. time horizon. Intuitively, the smaller the time horizon, the more important the temporal dynamics become when choosing the subset of most influential nodes of a given size. Fig. 4(a) shows the average total number of infected nodes against time horizon for a hierarchical Kronecker network with 1,024 nodes and approx. 2 edges per node. We consider a source set of cardinality $|A| = 10$ and we draw the transmission rate of each edge from a uniform distribution $\alpha \sim U(0, 5)$. The experimental results for all transmission rates configurations confirm the initial intuition, *i.e.*, the difference between INFLUMAX and other methods is greater for small time horizons.

Running time. Fig. 4(b) shows the average computation time per source added of our algorithm implemented (i) with lazy evaluation, (ii) with lazy evaluation and localized

source nodes with $m = 6$ hops and (iii) with lazy evaluation and limited transmission paths with $m = 6$ hops on a single CPU (2.3 Ghz Dual Core with 4 GB RAM). We use hierarchical Kronecker networks with an increasing number of nodes but approximately the same network density since real networks are usually sparse. Remarkably, the number of hops that we use in localized source nodes and limited transmission paths result in an approximation error for the influence $\sigma(A; T)$ of at most 10%, while achieving an speed-up of $\sim$5x for the largest network (2,048 nodes).

4.2. Experiments on real data

Experimental setup. We used the publicly available MemeTracker dataset, which contains more than 172 million news articles and blog posts (Leskovec et al., 2009). We trace the information in two different ways and then infer two different diffusion networks using NET-RATE (Gomez-Rodriguez et al., 2011).

First, we find more than 100,000 hyperlink cascades in the MemeTracker dataset. Each hyperlink cascade consists of a collection of time-stamped hyperlinks between sites (in blog posts) that refer to closely related pieces of information. From the hyperlink cascade data, we infer an underlying diffusion networks with the top (in terms of hyperlinks) 1,000 media sites and blogs. Second, we apply the MemeTracker methodology (Leskovec et al., 2009) to find 343 million short textual phrases. We cluster the phrases to aggregate different textual variants of the same phrase and consider the 12,000 largest clusters. Each phrase cluster is a MemeTracker cascade. Each cascade consists of a collection of time-stamps when sites (in blog posts) first mentioned any phrase in the cluster. From the Meme-Tracker cascades, we infer an underlying diffusion network with the top (in terms of phrases) 1,000 media sites and blogs. Then, we sparsify further the networks by keeping the 1,000 fastest edges since it has been shown that in the context of influence maximization, computations on sparsified models give up little accuracy, but improves scalability (Mathioudakis et al., 2011).

Solution quality. Fig. 5 shows the average total number of infected nodes against number of sources that INFLU-MAX achieves in comparison with other methods for both real networks, that were inferred from the hyperlink cascade and the MemeTracker cascade datasets, as described above. We set the time horizon to $T = 1.0$. Again, INFLU-MAX outperforms all other methods typically by $\sim$30%, by considering the temporal dynamics of the diffusion. Finally, we also compare INFLUMAX with the on-line bound from Th. 5 for the real network that we inferred from the hyperlink cascade dataset in Fig. 3(c). Similarly to the synthetic networks, the bound is not as tight as expected.

5. Conclusions

We have developed a method for influence maximization, INFLUMAX, that accounts for the temporal dynamics underlying diffusion processes. The method allows for variable transmission (influence) rates between nodes of a network, as found in real-world scenarios. Perhaps surprisingly, for the rather general case of continuous temporal dynamics with variable transmission rates, we can evaluate the influence of any set of source nodes in a network analytically using the work of Kulkarni (1986). In this analytical framework, we find the near-optimal set of nodes that maximizes influence by exploiting the submodularity of our objective function. In addition, the reevaluation of influence for changes on the network is straightforward and the algorithm parallelizes naturally by sink and source nodes.

We evaluated our algorithm on a wide range of synthetic diffusion networks with heterogeneous temporal dynamics which aim to mimic the structure of real-world social and information networks. Our algorithm is remarkably stable across different network topologies. It outperforms state of the art methods in terms of influence (*i.e.*, average number of infected nodes) for different network topologies, time horizons and source set sizes. INFLUMAX typically gives an influence gain of ∼25% and it achieves the greatest improvement for small time horizons; in such scenarios, the temporal dynamics play a dramatic role. We also evaluated INFLUMAX on two real diffusion networks that we inferred from the MemeTracker dataset using NETRATE. Again, it drastically outperformed the state of the art by ∼30%.

We believe that INFLUMAX provides a novel view of the influence maximization problem by accounting for the underlying temporal dynamics of diffusion networks.

References

Barabási, A.-L. and Albert, R. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

Bharathi, S., Kempe, D., and Salek, M. Competitive influence maximization in social networks. *Internet and Network Economics*, pp. 306–311, 2007.

Chen, W., Wang, Y., and Yang, S. Efficient influence maximization in social networks. In *KDD*, 2009.

Chen, W., Wang, C., and Wang, Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD*, 2010.

Clauset, A., Moore, C., and Newman, M. E. J. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

Erdős, P. and Rényi, A. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Science*, 5:17–67, 1960.

Georgiadis, L., Werneck, R.F., Tarjan, R.E., et al. Finding dominators in practice. *Journal of Graph Algorithms and Applications*, 10(1):69–94, 2006.

Gikhman, I.I. and Skorokhod, A.V. *The theory of stochastic processes*, volume 2. Springer Verlag, 2004.

Gomez-Rodriguez, M. and Schölkopf, B. Submodular Inference of Diffusion Networks from Multiple Trees. In *ICML*, 2012.

Gomez-Rodriguez, M., Leskovec, J., and Krause, A. Inferring Networks of Diffusion and Influence. In *KDD*, 2010.

Gomez-Rodriguez, M., Balduzzi, D., and Schölkopf, B. Uncovering the Temporal Dynamics of Diffusion Networks. In *ICML*, 2011.

Goyal, A., Bonchi, F., Lakshmanan, L.V.S., et al. Approximation Analysis of Influence Spread in Social Networks. *Arxiv preprint arXiv:1008.2005*, 2010.

Kempe, D., Kleinberg, J. M., and Tardos, É. Maximizing the spread of influence through a social network. In *KDD*, 2003.

Kulkarni, V.G. Shortest paths in networks with exponentially distributed arc lengths. *Networks*, 16(3):255–274, 1986.

Leskovec, J., Krause, A., Guestrin, C., et al. Cost-effective outbreak detection in networks. In *KDD*, 2007.

Leskovec, J., Backstrom, L., and Kleinberg, J. Meme-tracking and the dynamics of the news cycle. In *KDD*, 2009.

Leskovec, J., Chakrabarti, D., Kleinberg, J., et al. Kronecker graphs: An approach to modeling networks. *JMLR*, 11:985–1042, 2010.

Mathioudakis, M., Bonchi, F., Castillo, C., et al. Sparsification of influence networks. In *KDD*, 2011.

Nemhauser, GL, Wolsey, LA, and Fisher, ML. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1), 1978.

Provan, J.S. and Shier, D.R. A paradigm for listing (s, t)-cuts in graphs. *Algorithmica*, 15(4):351–372, 1996.

Richardson, M. and Domingos, P. Mining knowledge-sharing sites for viral marketing. In *KDD*, 2002.

Wallinga, J. and Teunis, P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, 160(6):509–516, 2004.