DIMACS Technical Report 2007-20
November 2007

Maximizing the Extent of Spread in a Dynamic Network

by

Habiba[1]
Department of Computer Science
University of Illinois at Chicago
Chicago, Illinois 60607

Tanya Y. Berger-Wolf[2]
Department of Computer Science
University of Illinois at Chicago
Chicago, Illinois 60607

[1]Work supported in part by the Fulbright Fellowship
[2]Work supported in part by the Microsoft award 14936

DIMACS is a collaborative project of Rutgers University, Princeton University, AT&T Labs–Research, Bell Labs, NEC Laboratories America and Telcordia Technologies, as well as affiliate members Avaya Labs, HP Labs, IBM Research, Microsoft Research, Stevens Institute of Technology, Georgia Institute of Technology and Rensselaer Polytechnic Institute. DIMACS was founded as an NSF Science and Technology Center.

ABSTRACT

Dynamic population phenomena, such as the spread of diseases, opinions, and behavior, can be modeled as processes that propagate in a network of interacting individuals. In this paper, we focus on the problem of identifying a set of individuals to initiate this spread so that the resulting extent of the spread is maximized. Kempe *et al.* [14] have solved this problem for the aggregate representation of the population network. Yet, real populations are inherently dynamic. In this paper we extend the approach of Kempe *et al.* to explicitly dynamic networks. We show that, for two common models of spread, maximizing the extent of spread in a dynamic network is an NP-hard problem. We present an $(1-1/e)$-approximation algorithm, evaluate the performance of the algorithm experimentally on real datasets and show that it performs well in practice. In addition, we compare the dynamic and the aggregate network representations both in terms of the resulting extent of spread and the actual set of initiating individuals that maximize the spread. We show that there are significant differences in both cases. Thus, we demonstrate that ignoring the dynamic aspects of data may result in extremely inaccurate answers and explicitly dynamic analysis is necessary.

1 Introduction

In this paper we study processes that spread through interactions in social networks that change over time. Diseases, ideas, fads, behavior– all are processes that spread in populations through individual contacts.

The idea of representing societies as networks of interacting individuals dates back to Lewin's earlier work of group behavior [16]. In this model individuals are represented as nodes and an edge exists if the two individuals have interacted at any time during the period they have been observed. This representation provides an aggregate view of the population. Spreading processes in such networks have been studied from different perspectives. Recently, as the data on very large social networks from blogs, knowledge-sharing sites, social-networking sites, newsgroups and chat rooms has become available, various aspects of social network analysis, including processes spreading over interactions in networks, have started to be considered from a computational perspective [14, 20, 21, 24, 25, 26, 27, 30].

Among the most important questions about a spreading process is its extent and how does that vary depending on the which set of individuals initiates the spread? These questions have been considered from different perspectives in the domains of epidemiology [4, 6, 19, 22, 24, 25], diffusion of technological innovations and adoption of new products [1, 2, 5, 11, 15, 27], as well as voting, strikes, rumors [12, 21, 31]. Kempe *et al.* in [14] have algorithmically addressed the problem of finding the most influential individuals in the aggregate network model.

However, societies are inherently dynamic, that is, both the individuals making up the society and the interactions among them change with time. Moreover, there are processes that spread only over those interactions that exists at a given moment. For example, in the context of disease spread, two individuals may be in contact for a long period of time while they are in the susceptible stage, but they may not be in close proximity of each other when either of them gets infected and thus can infect the other. Here, the explicit concurrency and order of interactions affects the resulting spread of the disease. While for many diseases we do not have the exact data on interactions among individual entities, for diseases like bird flu we have the exact sequence of contacts among commercial birds and, thus, can it to predict the spread. Zebra behavior is another such example where the leaderships in herds and harems is based on which animals are within the observable distance of the other. Zebras are influenced by those with whom they have been in close proximity most recently. Thus, a change in the behavior in zebra population will spread only over those contacts that have occurred within a short period of time. Here again, we have the exact observed sequence of movements (for example, from GPS collars) and our conclusions can be tested by comparing those with the movements observed in the field. In both examples, the sequence of time is critical. For all such spreading processes the aggregate model provides an inaccurate estimate both for the extent of spread and for finding influential individuals.

Using explicitly dynamic networks framework we provide an approximate solution to the problem of identifying the most influential individuals in a population. We build on ideas presented in [14], where it has been shown that finding the minimum set of most influential individuals is NP-hard in aggregate social networks.

We show that the problem of finding the optimal set of influential individuals remains NP-hard for dynamic social networks. We provide a $(1 - 1/e)$-approximation algorithm for finding a set of most influential individuals (Section 4). We compare the optimal solution with our proposed approximate algorithm using real data (Section 6.2) and show that our approximation algorithm performs well in practice.

We use two of the most common transmission models: the Independent Cascade model [10, 11] and the Linear Threshold model [13] for the computational analysis of the spreading process. We compare the resulting spreads, quantitatively and qualitatively, for both the aggregate and the dynamic networks. We experimentally show that for Independent Cascade model, the extent of spread in the aggregate network is an overestimate of the actual spread taking place in dynamic networks. Whereas, for the Linear Threshold model, the aggregate network underestimates the actual spread measured in dynamic networks. Moreover, we show that the set of influential individuals derived from the aggregate network model may be completely different from the one obtained from the explicitly dynamic analysis. Hence, we computationally show that ignoring the time order of interactions in general gives results that do not relate to the actual spreading process taking place in the population. Algorithms that explicitly address the dynamic aspects of the data are necessary and must be used where the data on the timeline of interactions is available and we provide such an algorithm.

2 Definitions and Preliminaries

2.1 Dynamic Network

Traditionally, a social network is modeled as a graph with nodes representing individual entities and edges corresponding to interactions between those individuals. In a standard aggregate representation of a social network, an edge exists between a pair of individuals if they have interacted at any point during the period of data collection. However, such aggregate view discards the temporal information about the timing and order of interactions. Instead, we use an explicitly dynamic model of a social network.

A dynamic network is a time series of static social networks. The time is divided into discrete steps and all interactions that occur from one timestep to another are considered to have occurred in that timestep and not in between. Each static network is a snapshot of individuals and their interactions observed at that timestep. The dynamic network then describes the evolution of interactions among individuals over a period of time. Figure 1 gives an example of different dynamic networks and the same corresponding aggregate network representation. We now give a formal definition of a dynamic network.

Definition 1 *Let $\{1, \ldots, T\}$ be a finite set of discrete timesteps. Let $I = \{1, \ldots, n\}$ be a set of individuals. Let $G_t = (V_t, E_t)$ be a graph representing the snapshot of a static network at time t. $V_t \subseteq I$ present at time t and $(u_t, v_t) \in E_t$ if the individuals u and v interacted at time t.*

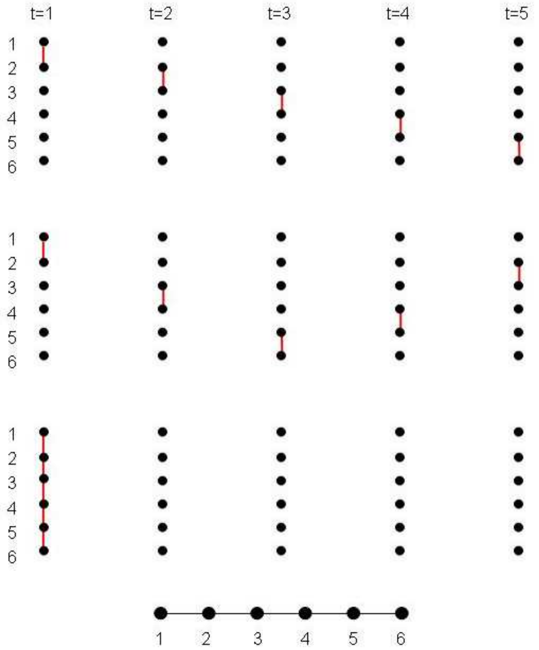

Figure 1: 3 dynamic (top) and corresponding 1 aggregate (bottom) social networks of 6 individuals interacting over 5 timesteps.

A dynamic network $G = \langle G_1, \ldots, G_T \rangle$ *is the graph* $G = (V, E)$ *of the time series of graphs* G_t *such that* $V = \bigcup_t V_t$ *and* $E = \bigcup_t E_t \cup \bigcup_{V,t} (v_t, v_{t+1})$.

In general, in a dynamic network both edges and individuals can change over time. However, for simplicity, we assume that the set of individuals remains the same for the entire time period: $I = V_1 = \ldots = V_T$. Nonetheless, our results easily extend to networks where this assumption does not hold. The edges E_t may be directed or undirected and weighted or unweighted. The self-edges of the type (v_t, v_{t+1}) are always directed.

2.2 Models of Spread

A spreading phenomenon in a network can be described formally using many models of transmission. For this paper we use two models that have been extensively studied in the context of social networks, viral marketing, and epidemiology [10, 11, 12, 18, 22]:

Independent Cascade: The independent cascade model was first introduced in [10, 11] in the context of word-of-mouth marketing. This is also the most commonly used simple model to study disease transmission in networks [4, 19, 22, 24, 25]. In this model, transmission from one individual to another happens independently of interactions with all the other individuals.

Linear Threshold: Models of collective behavior, are developed for situations where individuals choose between alternatives based on how many other individuals connected to them choose those alternatives. The key concept is that of a “threshold” which is the cumulative number of neighbors of an individual that must make a decision before the individual does so, too [12].

We now describe both models in more detail.

2.2.1 Independent Cascade Model.

The independent cascade model describes a spreading process comprising of two sets of individuals, active and inactive. The process unfolds in discreet timesteps. In each timestep, each active individual attempts to activate each of its neighbor independently. The activation of each inactive neighbor is determined by a probability of success. If an active individual succeeds in affecting any of its neighbors, those neighbors become active in the next time step. Each attempt of activation is independent of all previous attempts as well as the attempts of any other individual to activate a common neighbor.

More formally, the input to the independent cascade model is a set $A_0 \subseteq I$ of active individuals. In a dynamic network, an active individual v_t in timestep t tries to activate each of its currently inactive neighbors w_t only once with a probability p_{v_t,w_t}, independent of all the other neighbors. If v_t succeeds in activating w_t at timestep t, then w_t will be active in step $t+1$, whether or not $(w_{t+1}, v_{t+1}) \in E_{t+1}$. If v_t fails in activating w_t, and at any subsequent timestep v_{t+i} gets reconnected to w_{t+i}, it will again try to activate w_{t+i}. The process runs for a finite number of timesteps T or until no more activations are possible. The process has two outputs: the set A_T of active individuals at the end of the spreading process and the size $|A_T|$ of the resulting set of active individuals. We denote by $\sigma(A_0) = A_T$ the correspondence between the initial set A_0 and the resulting set of active individuals A_T.

The spreading process in the independent cascade model in a dynamic network is different from the aggregate network in one important aspect. In the aggregate case, each individual v uses all its attempts of activating each of its inactive neighbors w with the same probability $p_{v,w}$ in one timestep t. This is the timestep right after the individual v itself becomes active. After that single attempt the active individual becomes latent: that is, it is active but unable to activate others. However in the dynamic network model as defined above, the active individuals never become latent during the spreading process. Note that the aggregate model is always equivalent to a dynamic model defined by the order of the activation of the nodes.

For this paper, we only consider the progressive case in which an individual converts from inactive to active but never reverses.

2.2.2 Linear Threshold Model.

The linear threshold model also describes the spread over two sets of individuals, active and inactive. Each inactive individual has a certain susceptibility to become active, which is denoted by the individual's "threshold". Each active individual has a certain "weight" of influence over each of its inactive neighbors. An individual becomes active if the accumulated weight of all its active neighbors exceeds the individual's susceptibility threshold.

More formally, the linear threshold model is defined by two parameters. For each individual v, a threshold $\theta_v \leq 1$ indicates the latent tendency of this individual v to be activated. For each edge $(w_t, v_t) \in E_t$ (for $0 \leq t \leq T$) the weight b_{w_t,v_t} is the influence of the individual w on v, that is, w's ability to activate v. For each v, $\sum b_{w,v} \leq 1$.

The spreading process described by the linear threshold model in a dynamic network starts with a given set of thresholds θ_v assigned to each individual v. The initial set of active individuals is A_0. The process unfolds in discreet timesteps, $1 \ldots T$. At each step t, each inactive individual v_t is influenced by the set of its active neighbors. The inactive individual v_t becomes active at timestep $t+1$, if $\sum b_{w_t,v_t} \geq \theta_v$. If $\sum b_{w_t,v_t} < \theta_v$ then v remains inactive and at every subsequent timestep, a new attempt is made to activate it. Each attempt is independent of any previously made attempts. The process continues for T steps or until no more activations are possible. The outputs, again, are the set of individuals active at time T, A_T and the size of that set $|A_T|$. We denote by $\sigma(A_0) = A_T$ the correspondence between the initial set A_0 and the resulting set of active individuals A_T.

The spreading process in the dynamic network is different from the aggregate network, where each active individual is given only one attempt to activate any of its inactive neighbor. Again, each spreading scenario in the aggregate network is equivalent to the spread in some realization of a dynamic network.

The difference between independent cascade and the linear threshold models is that each attempt of activation is independent of the attempts by all the other active individuals while in the later model each inactive individual is influenced by the aggregated weight of all its active neighbors.

3 Problem Statement

We now formally define the Dynamic Influence Maximization problem:

Given: a dynamic social network $G = \langle G_1, \ldots, G_T \rangle$ over T timesteps and an integer k

Find: the initial set of individuals A_0 of size $|A_0| = k$

Objective: maximize the extent of spread initiated by A_0 as measured by the expected size of the set of active individuals $|\sigma(A_0)| = |A_T|$ in G after T timesteps.

We consider this problem for both spreading models described above. We show that the Dynamic Influence Maximization problem is NP-hard for both models but it can be efficiently approximated using a simple greedy approach.

4 Complexity

Kempe *et al.* [14] show that the Influence Maximization problem in the aggregate networks is NP-hard using reductions from the Set Cover problem for the Independent

Cascade model and from the VERTEX COVER problem for the Linear Threshold model. Both reductions are to instances of the INFLUENCE MAXIMIZATION problem where the spread happens in one timestep. Thus, those aggregate instances of the INFLUENCE MAXIMIZATION problem are special cases of the DYNAMIC INFLUENCE MAXIMIZATION problem with $T = 1$ timestep. Therefore, the DYNAMIC INFLUENCE MAXIMIZATION problem for both spread models remains NP-hard.

5 Approximation Algorithm

5.1 The Overall Approach

We now present an approximation algorithm for both models of spread. We use the notion of *submodular functions.*

5.1.1 Submodular functions.

Consider an arbitrary function $f(.)$ that maps subsets of a finite set U to non-negative real numbers.

Definition 2 *A function $f : [U] \to R^+$ is* submodular *if it satisfies the natural diminishing returns property. This property states that the marginal gain of adding an element v to a set S is at least as great as adding v to a set T, where $S \subseteq T$:*

$$\forall v,\ \ S \subseteq T:\ \ f(S \cup v) - f(S) \geq f(T \cup v) - f(T)^1.$$

One of the basic properties of submodular functions is that if a non-negative function f is submodular then it is monotonically increasing, i.e.

$$f(S \cup v) \geq f(S).$$

For the DYNAMIC INFLUENCE MAXIMIZATION problem we find a set S of individuals such that $f(S)$ is maximal, that is, adding any other individual to S will not increase the value of $f(S)$. This is an NP-hard problem for any submodular function by reduction from the HITTING SET problem [3, 23]. A simple greedy hill climbing algorithm [23] gives a $(1 - 1/e)$-approximation of the optimal solution. We use a similar greedy approach to achieve a $(1 - 1/e)$-approximation and show that you cannot do any better for the specific case of the DYNAMIC INFLUENCE MAXIMIZATION problem.

5.1.2 The algorithm.

The general approximation algorithm for the DYNAMIC INFLUENCE MAXIMIZATION problem is as follows:

[1] We abuse the set notation slightly by denoting the singleton set $\{v\}$ as just v for simplicity.

Let f be a non-negative (monotonic) submodular function. Let S be the set of size k obtained by selecting one element at a time, each time choosing an element that provides the largest marginal increase in the function's value. Let the optimal solution be S^*, which is the set that maximizes the value of f over all k-elements sets. Then $f(S) \geq (1 - 1/e)f(S^*)$. That is, S guarantees $(1 - 1/e)$-approximation.

The definition of the approximation ratio assumes that the function f is either already known or is easily computable. However, for the stochastic spreading process there is no such known function. The best that can be done is to simulate the spreading process on sample inputs and retrieve the results, which are the sets of active individuals A_T at time T. Based on these results we can compute a reasonably accurate estimate of the function f. Kempe *et al.* [14] show that for some $\epsilon > 0$, there is a $\gamma > 0$ such that using $(1+\gamma)$-accurate estimates of f results in a $(1 - 1/(e - \epsilon))$-approximation of the aggregate Influence Maximization problem. This provides an approximation scheme that achieves an approximation ratio arbitrarily close to the $1 - 1/e$. We use this approximation approach for the Dynamic Influence Maximization problem with the two spreading models. We prove that the influence maximization function for the two models is submodular, the rest of the proof follows as described.

5.2 The Independent Cascade Model

Using the above approach we prove that for an arbitrary instance of the independent cascade model, the resulting influence function $\sigma(.)$ is submodular. That is, the increase in the expected number of activated individuals for any set $S \subseteq T$ and an individual v_i is $\sigma(S \cup v_i) - \sigma(v_i) \geq \sigma(T \cup v_i) - \sigma(v_i)$. Unlike the aggregate case described in [14], the dynamic network is well defined by ordering of individual to individual connections based on the timesteps. For each timestep t, to find whether an individual v_t is successful in activating an individual w_t we flip a coin for the edge (v_t, w_t) with the bias p_{v_t,w_t}. The probability $p_{v,v}$ for the edges of the form (v_t, v_{t+1}) for all timesteps t and all individuals is set to 1. The coin flips can be performed earlier than the processing of a given edge. Thus, we flip all the coins for G at the beginning of the spreading process and use them when we process the corresponding edge. We call the edges for which the coin flips indicate success "live" and the rest are "blocked". All edges of type (v_t, v_{t+1}) have probability 1 of activation, therefore, they are live by default. Note, that an individual u will be active if only if there is a path from some individual in A_0 to u consisting entirely of live edges. By fixing the outcomes of coin flips and selecting a set A_0 at the beginning, we can easily find the set of active individuals at the end of the spreading process by tracing a path of live edges from each individual in A_0.

Let X be a sample set of outcomes of coin flips for all edges. We define the activation outcome of an active individual v by $R(v, X)$, which is the set of individuals activated by v. Given a sample set of coin flips X and an initial set A_0, we denote the number of active individuals at the end of the spreading process by $\sigma_X(A_0)$. Note, that $\sigma_X(A_0)$ is the size of the union of all $R(v, X)$ where $v \in A_0$. The overall (incomputable) influence function $\sigma(A_0)$

is the expectation of $\sigma_X(A_0)$ over all possible sets of coin flips X. We show that $\sigma_X(A_0)$ is submodular.

Lemma 1 *$\sigma_X(A_0) = \bigcup_{v \in A_0} R(v, X)$ is a submodular function.*

Proof 1 *Let S and T be two sets such that $S \subseteq T$. Consider the quantity $\sigma_X(S \cup v) - \sigma_X(S)$. By definition*

$$
\begin{aligned}
\sigma_X(S \cup v) \quad &- \quad \sigma_X(S) \\
&= \quad |\bigcup_{u \in S \cup v} R(u, X) - \bigcup_{u \in S} R(u, X)| \\
&= \quad |R(v, X) - \bigcup_{u \in S} R(u, X)| \\
&\geq \quad |R(v, X) - \bigcup_{u \in T} R(u, X)| \\
&= \quad \sigma_X(T \cup v) - \sigma_X(T)
\end{aligned}
$$

That is, the number of elements in $R(v, X)$ that are not already in $\cup_{u \in S} R(u, X)$ is at least as big as the number of elements not in the larger set $\cup_{u \in T} R(u, X)$. This inequality shows that $\sigma_X(A_0)$ is submodular for the independent cascade model.

5.3 The Linear Threshold Model

For the linear threshold model as well, we show that for an arbitrary instance of the linear threshold model, the resulting spreading function $\sigma(.)$ is submodular. Unlike the independent cascade model, a general influence function for the linear threshold model based on fixed choices of thresholds θ_v for each individual v is not submodular. This is due to the *accumulated* effect of individuals on other individuals. As defined above, each individual u_t has an influence $b_{v_t,u_t} \geq 0$ on each of its neighbor v_t, such that the total influence on v_t at time t is $\sum_{(u_t,v_t) \in E_t} b_{v_t,u_t}$. We convert the linear threshold spreading process to another form. We use the approach of [14], to resolve the non-submodularity of the spreading function. At time t, for each individual v_t we pick at most one edge incident on v_t from any neighbor individual u_t with probability b_{v_t,u_t}, and do not pick any other incident edge with the probability $1 - \sum_{(u_t,v_t) \in E_t} b_{v_t,u_t}$. If u_t is not connected to v_t then $b_{v_t,u_t} = 0$. An edge (u_t, v_t) is called "live" if it has been picked for activation and all other edges connected to v_t are called "blocked". Note, that unlike the independent cascade model, where each edge is live independent of the other edges and an individual can have multiple live edge paths, in this process, each active individual belongs to a unique live edge path.

In this modified version of the linear threshold model, the submodularity property of the function $\sigma(.)$ holds as in the independent cascade model. We pick a sample distribution X of live/blocked edges, and define $R(v, X)$ as the set of individuals reachable through live edge paths from v. Thus, for the initial set A_0, it follows that $\sigma_X(A_0)$ is the cardinality of

$\cup_{v \in A_0} R(v, X)$, which is the submodular function of A_0. Finally, since $\sigma(.)$ is an expectation of $\sigma_X(.)$ over all X, it is a non-negative linear combination of the functions $\sigma_X(.)$ and hence is also submodular.

Now, we show that the original linear threshold model is equivalent to the modified version.

Lemma 2 *The distribution over active sets obtained by running the linear threshold process to completion starting from A_0, is equivalent to the distribution of active sets reachable by A_0 under the random selection of live edges defined above.*

Proof 2 *Consider the active set A_t at time t, and let v be an individual which is not yet active. This individual becomes active at $t+1$ if the set of individuals $A_t - A_{t-1}$ pushes it over the threshold θ_v. Thus, the probability of activation of each individual can be defined as:*

$$Pr(activation\ of\ v) = \frac{\sum_{u \in A_t - A_{t-1}} b_{v_t, u_t}}{1 - \sum_{u \in A_{t-1}} b_{v_{t-1}, u_{t-1}}}.$$

Similarly, for the redefined linear threshold model using random live-edge pattern, we run the spreading process by revealing the identities of live edges gradually as follows. At time t, A_t is the set of active individuals. We pick an individual v_t outside of A_t and check whether v_t has a live edge connecting it to any individual in A_t. If yes, then v_t becomes active in time $t+1$. The activation of v_t at $t+1$ shows that the live edge of v_t comes from the set $A_t - A_{t-1}$. This is equivalent to the probability

$$Pr(activation\ of\ v) = \frac{\sum_{u \in A_t - A_{t-1}} b_{v_t, u_t}}{1 - \sum_{u \in A_{t-1}} b_{v_{t-1}, u_{t-1}}}.$$

Thus, the probability of activation through the live edge path is equivalent to the probability of activation of an individual obtained by the original spreading process. Hence, by induction over the time steps, we can show that the distribution over active sets obtained by linear threshold model is the same as that of the live edge path process.

This completes the proof of the submodularity of $\sigma(.)$ for both models of spread. Thus, as outlined in Section 5.1.2, a simple greedy heuristic of selecting one by one, the initial individuals that provide the largest marginal gain, gives a $(1 - 1/e)$-approximation. This is a theoretical bound. We now show experimental results of evaluating the performance of the approximation algorithm in practice, as well as compare the aggregate and the dynamic influence maximization problem formulations and solutions.

6 Experiments

6.1 Types of experiments

We have performed three sets of experiments for each of the two models of spread.

1. To evaluate the performance of the approximation algorithm in practice, we compare the optimal solution (found by exhaustive search) to the results of the greedy approximation algorithm for the dynamic network.
2. To evaluate the differences in the estimates of the extent of spread in aggregate and dynamic networks, we quantitatively compare the extent of spread given by the approximation algorithm in both networks.
3. To evaluate the differences in the identification of the influential individuals in aggregate and dynamic networks, we qualitatively compare the extent of spread in the dynamic network using the initial set obtained from the approximation algorithm run on the aggregate network.

We now describe the experiments in more detail.

6.1.1 Optimal vs Approximate:

For small datasets we simulate the spread using the two transmission models. We compare the extent of spread resulting from the optimal initial set of active individuals to that from the set found by the approximation algorithm. We use small datasets so that we can compute the optimal solution using the exhaustive search. We compare the difference in the extent of spread for the initial sets sizes of 3, 5, and at most 10 individuals. Calculating the optimal solution for the initial set sizes of more than 10 individuals is not feasible using the exhaustive search.

Let k be the size of the set of the initial individuals we are interested in identifying. For the optimal solution we use the following algorithm for each transmission model.

For each possible subset of k individuals:

1. *Simulate the spreading process for* 1000 *trials.*
2. *Mark each individual not in the initial set as activated if that individual was active in more than half of the trials.*
3. *Pick the initial set that results in the maximum number of active individuals. This is the optimal set that maximizes the spread.*

For the approximation algorithm, we use the following greedy approach with each transmission model.

1. *For each individual calculate the extent of spread when only that individual was in the initial set. Run each experiment for* 1000 *trials for each individual.*
2. *Calculate the average number of individuals activated by the singleton initial set at the end of the* 1000 *trials.*
3. *Pick the individual that gives the maximum average number of the activated individuals.*

4. *Remove that individual and all those individuals that were activated by it in more than half of the trials from the population.*

5. *Repeat steps (1–4) until k most influential individuals have been picked.*

6.1.2 Dynamic vs Aggregate - Quantitative:

For the second set of experiments we compare the extent spread in the aggregate and the dynamic networks. We use the same approximation algorithm described in the previous set of experiments. In the aggregate network, the edges between individuals at different timesteps are represented as multiple edges. The spreading process in the aggregate network proceeds as described in Section 2.2.

6.1.3 Dynamic vs Aggregate - Qualitative:

For the third set of experiments we are interested in qualitative comparison of aggregate and dynamic networks. We use identical initial sets for the two networks. We compare the sets of activated individuals in each network. The initial sets used in both cases are the ones obtained by the approximation algorithm run on the aggregate networks.

6.2 Data sets

The following are the datasets used in the experiments.

6.2.1 Grevy:

Populations of Grevy's zebras (*Equus grevyi*) were observed by biologists [8, 9, 28, 29] over a period of June–August 2002 in the Laikipia region of Kenya. Predetermined census loops were driven on a regular basis (approximately twice per week) and individuals were identified by unique stripe patterns. Upon sighting, an individual's GPS location was taken. In the resulting dynamic network, each node represents an individual animal and two animals are interacting if their GPS locations are the same. The dataset contains 28 individuals interacting over a period of 44 timesteps.

6.2.2 Onagers:

Populations of wild asses (*Equus hemionus*), also known as onagers, were observed by biologists [28, 29] in the Little Rann of Kutch, a desert in Gujarat, India, during January–May 2003. This data is also obtained from visual scans, as in Grevy's zebra case. The dataset contains 29 individuals over 82 timesteps.

6.2.3 Portland:

The Portland dataset [7] is a statistically correct simulation of the movement of 1.6 million individuals of Portland on a daily basis. Each node in the network represents an individual and the two individuals are interacting if they are at the same "location", as defined by the simulation. We use a subset of 1588 individuals over 80 timesteps. This is the largest sample data set we used for the experiments. This network is further divided into two large components. For some experiments we use the smaller of the two components of 211 individuals. We use this dataset to compare aggregate and dynamic networks.

6.2.4 DBLP:

This data set is a sample of the *Digital Bibliography and Library Project* [17]. This is a bibliographic dataset of publications in Computer Science. We use a cleaned version of the data from 1967–2005. In the dynamic network each node represents an individual author and two authors are interacting if they are co-authors on a paper. A year is one timestep. The sample we used contains 1374 individuals and 38 timesteps. We use this dataset to compare aggregate and dynamic networks.

7 Experimental results

7.1 Independent Cascade

7.1.1 Optimal vs Approximate.

For this set of experiments we use small data sets, Grevy's and Onagers, with less than 30 individuals each. We run the experiments for the optimal and the approximate (greedy) initial sets. We compare the spread for the initial set sizes of 3 to at most 10 individuals. The probability of activation was set the same for all the edges. We ran the experiments with a range of probability values. For $pr > 0.2$ the entire population is activated quickly, while for $pr = 0.05$ only a small set of individuals is activated at the end. Here we show the results for $pr = 0.2$.

The first set of simulations is on the Grevy's dataset. We compare the results of initial set sizes of 3, 5, 7 and 10 individuals. We could not get the optimal solution for the initial set size of 10 since the number of subsets of size 10 for the dataset, $\binom{28}{10} > 2.8^{10}$, is too big and the exhaustive search is not feasible. However, the value of the optimal solution for 10 initial individuals is at least 27 (which is the value of the optimal solution for 7 initial individuals) and is more likely to be 28, the entire population.

The second set of simulations was conducted on the Onagers data. Initial set sizes are 3, 5, and 6. We did not need to consider larger initial sets since 6 initial individuals activate the entire population. The results for both datasets are shown in Table 1.

The experiments show that the results obtained using the approximation algorithm are within 80% accuracy of the optimal solution. This is better than the $(1 - 1/e) \approx 0.632$

Table 1: Optimal Vs Approximate(Independent Cascade, Pr=0.2)

Data Set	Pop.size	Init set size	Opt	Approx
Grevy's	28	3	20	18
		5	23	20
		7	27	22
		10	27–28	25
Onagers	29	3	26	26
		5	28	28
		6	29	29

accuracy predicted by the theoretical analysis.

7.1.2 Dynamic vs Aggregate: Quantitative Comparison.

In the second set of experiments, we compare the spread in the dynamic and the aggregate networks achieved by the approximation algorithm. We use the large data sets for which finding the optimal solution is not feasible: the 211 individuals component of the Portland dataset and the DBLP dataset. The initial sets are of the sizes 3, 5, and 10 individuals. The probability of activation is set to 0.2, the same for all edges. The results for the Portland and DBLP datasets are given in Table 2. We observed that the extent of spreads resulting from the same sizes of the initial sets of individuals are much greater in the aggregate networks than in the dynamic networks. The difference is much more pronounced in the DBLP dataset, the sparser of the two.

Table 2: Dynamic vs Aggregate: Quantitative(Independent Cascade, Pr= 0.2)

Data set	Pop.size	Init set	Aggregate	Dynamic
Portland	211	3	186	96
		5	211	125
		10	211	181
DBLP	1374	3	58	7
		5	65	11
		10	76	22

7.1.3 Dynamic vs Aggregate: Qualitative Comparison.

In the last set of experiments we compare the extent of spreads in the dynamic networks obtained by using the initial sets resulting from running the approximation algorithm on the aggregate networks. We use the Portland and the DBLP datasets. The initial set sizes are

3 and 5 for Portland, since 5 initial individuals activate the entire network, and 3, 5, and 10 for DBLP. The probability of activation is set to 0.2. The results are shown in Table 3.

Table 3: Dynamic vs Aggregate: Quantitative(Independent Cascade, Pr= 0.2)

Data set	Pop.size	Initial set	Aggregate	Dynamic
Portland	211	3	186	40
		5	211	65
DBLP	1374	3	58	4
		5	65	9
		10	76	18

We take the best (approximate) set obtained in the aggregate networks and use it to initiate the spread in the dynamic networks. However, the resulting extent of spread in the dynamic networks is significantly less than the extent of spread obtained using the approximation algorithm directly on the dynamic networks. This experiments demonstrates that the optimal set of initial individuals is not the same in the aggregate and the dynamic networks. Moreover, a set that performs well in the aggregate network may perform very poorly in the dynamic network.

7.2 Linear Threshold

We performed the same sets of the experiments using the linear threshold transmission model. For all the experiments we used initial sets of sizes up to 10 or the size that resulted in the entire population being activated. The threshold probabilities were set randomly for each individual.

7.2.1 Optimal vs Approximate.

We calculate the extent of spread using the optimal exhaustive search and the greedy approximation algorithm on the Grevy's and the Onagers datasets. The results are show in Table 4. The results demonstrate that the approximate algorithm is 100% accurate on these datasets for the linear threshold model and, thus, can be used in practice for a good estimate of the extent of spread.

Using the same data as in the experiments for the independent cascade transmission model, we observe that the extent of spread in linear threshold model is greater for the same size of the initial set. For Grevy's, population size of 28, 6 initial individuals activate the entire population. For Onagers, with the population size of 29, the initial set of size 3 activateds the entire population, for both the optimal and the approximate solutions.

The experiments show that for the linear threshold model the approximate algorithm performs very well in practice. In fact, it is as good as the optimal solution for the datasets considered in the experiments.

Table 4: Optimal Vs Approximate(Linear Threshold)

Data Set	Pop.size	Init set	Opt	Approx
Grevy's	28	3	25	25
		5	27	27
		6	28	28
Onagers	29	3	29	29

7.2.2 Dynamic vs Aggregate: Quantitative Comparison.

To compare the extent of spread obtained by the approximation algorithm in the dynamic and the aggregate networks, we use the complete Portland data set of 1588 individuals and DBLP data set. The initial set sizes were set to 3, 5, and 10 individuals. None of the initial sets resulted in activation of the entire population but the 10 individuals' set came very close to it. The results are shown in Table 5. As in the independent cascade case, the extent of spread in the dynamic and the aggregate networks are very different. However, unlike in the independent cascade, here the extent of spreads in the dynamic networks is greater than in the aggregate networks.

Table 5: Dynamic vs Aggregate: Quantitative(Linear Threshold)

Data set	Pop.size	Init set	Aggregate	Dynamic
Portland	1588	3	665	1569
		5	724	1571
		10	1396	1576
DBLP	1374	3	961	1165
		5	971	1167
		10	993	1172

7.2.3 Dynamic vs Aggregate: Qualitative comparison.

In the last set of experiments with the linear threshold transmission model we compare the extent of spread in dynamic networks obtained by using the initial sets resulting from running the approximation algorithm on aggregate networks. Again, we use the Portland and the DBLP datasets with the initial set sizes 3, 5, and 10 individuals. The results are shown in Table 6.

As in the independent cascade case, the best initial set obtained for aggregate networks perform extremely poorly on dynamic networks. Thus, also for the linear threshold model, to find the right set of initial individuals we must perform the analysis on the dynamic networks directly.

Table 6: Dynamic vs Aggregate: Qualitative(Linear Threshold)

Data set	Pop.size	Init Set	Aggregate	Dynamic
Portland	1588	3	665	25
		5	724	30
		10	1396	71
DBLP	1374	3	961	62
		5	971	70
		10	993	84

8 Conclusions and Future Work

Processes spreading in social networks have been studied in many different contexts and with various objectives. Yet, most of the analysis techniques are based on the aggregate model of a social network, with the information about the concurrency and the order of interactions discarded. However, for some processes, such as the spread of a communicable disease or behavior in animal populations, the outcome depends on the explicit timing of interactions. Moreover, the data on the timeline of those contacts is increasingly becoming available. Thus, it is now necessary and possible to address the question of how much and under what circumstances does the dynamic nature of social interactions affect the outcomes of processes spreading through those interactions.

In this paper we define the problem of finding the set of individuals that maximizes the extent of a spreading processes in explicitly dynamic social networks. We show that this Dynamic Influence Maximization is NP-hard for two common models of spreading processes. We propose a $(1 - 1/e)$-approximation algorithm that performs well in practice.

We compare the results of Influence Maximization [14] in the dynamic and the aggregate social networks. The extent of spread in the two types of representations is different and the difference depends on the models of spread and the dynamic topology of the networks. Moreover, the individuals that maximize the extent of spread in the dynamic and aggregate graphs are different.

Thus, the aggregate representation not only gives an inaccurate information about networks, using it to make predictions about the spread of ideas, behavior, or diseases in a social network may lead to incorrect conclusions. Hence, it is crucial to use dynamic networks where the interactions among individuals are sensitive to the time order.

However, in many instances, we must provide an estimate of the extent of spread and some idea of the individuals that might be its facilitators even when the explicit dynamic information about the interactions is not available. Thus, it is important to identify some measures that give theoretical estimates of the spread in dynamic networks based on aggregate graphs. Given an aggregate representation of a network, we can then analytically estimate the worst and the best case dynamic scenarios that correspond to that representation. There may also be some topological indicators, such a dynamic version of the centrality of an individual, that help identify influential individuals.

There are many other possible extensions of our work. In this paper we focused on identifying the facilitators of a spreading process. A very important but computationally a much harder problem is to identify the set of individuals that effectively prevent a process from spreading in a network. Such individuals would be the best candidates for vaccination in case of a spread of a disease. In this paper, we simplified the models by considering only the progressive case, where an individual remains active once it has been activated. This work naturally extends to non-progressive models. Finally, models of spread other than the two discussed in this paper can be used to get a better comparison between the dynamic and the aggregate networks.

In conclusion, understanding how processes spread in networks is an important and growing area within many application domains. We must find the right models, measures, and algorithms to study explicitly dynamic networks. Our work demonstrates the importance of dynamic information and takes the first steps in analyzing the behavior of spreading processes in explicitly dynamic data.

References

[1] K. Carley. Communicating new ideas: The potential impact of information and telecommunication technology. *Technology in Society*, 18(2):219–230, 1996.

[2] L. Chen and K. Carley. The impact of social networks in the propagation of computer viruses and countermeasures. *IEEE Trasactions on Systems, Man and Cybernetics*, forthcoming.

[3] G. Cornuejols, M. Fisher, and G. Nemhauser. Location of bank accounts to optimize float. *Management Science*, 23, 1977.

[4] Zoltán Dezsö and Albert-László Barabási. Halting viruses in scale-free networks. *Physical Review E*, 65(055103(R)), 2002. DOI: 10.1103/PhysRevE.65.055103.

[5] P. Domingos. Mining social networks for viral marketing. *IEEE Intelligent Systems*, 20:80–82, 2005.

[6] S. Eubank, H. Guclu, V.S. Kumar, M.V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429:429:180–184., Nov 2004. Supplement material.

[7] S. Eubank, V.S.A. Kumar, M.V. Marathe, A. Srinivasan, , and N. Wang. Structural and algorithmic aspects of massive social networks. In *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 718–727, 2004.

[8] I. R. Fischhoff, S. R. Sundaresan, J. Cordingley, H. M. Larkin, M-J. Sellier, and D. I. Rubenstein. Social relationships and reproductive state influence leadership roles in movements of plains zebra (*equus burchellii*). *Animal Behaviour*, 2006. Submitted.

[9] I. R. Fischhoff, S. R. Sundaresan, J. Cordingley, and D. I. Rubenstein. Habitat use and movements of plains zebra (*equus burchelli*) in response to predation danger from lions. Submitted.

[10] J. Goldenberg, B.Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, 2001.

[11] J. Goldenberg, B. Libai, and E. Muller. Using complex systems analysis to advance marketing theory development. *Academy of Marketing Science Review*, 2001.

[12] M. Granovetter. Threshold model for collective behavior. *American Journal of Sociology*, 83:1420–1443, 1978.

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

[14] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[15] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. In *EC '06: Proceedings of the 7th ACM conference on Electronic commerce*, pages 228–237, New York, NY, USA, 2006. ACM Press.

[16] K. Lewin. *Principles of Topological Psychology.* New York: McGraw Hill, 1936.

[17] Michael Ley. Digital bibliography & library project (DBLP). `http://dblp.uni-trier.de/`, December 2005. A digital copy of the databse has been provided by the author.

[18] M. Macy. Chains of cooperation.:threshold effects in colllective action. *American Sociological Review*, 56, 1991.

[19] Robert M. May and Alun L. Lloyd. Infection dynamics on scale-free networks. *Physical Review E*, 64(066112), 2001. DOI: 10.1103/PhysRevE.64.066112.

[20] L. A. Meyers, M.E.J. Newman, and B. Pourbohloul. Predicting epidemics on directed contact networks. *Journal of Theoretical Biology*, 240:400–418, 2006.

[21] Yamir Moreno, Maziar Nekovee, and Amalio F. Pacheco. Dynamics of rumor spreading in complex networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 69(6):066130, 2004.

[22] M. Morris. Epidemiology and social networks:modeling structured diffusion. *Sociological Methods and Research*, 22(1):99–126, 1993.

[23] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. *Management Science*, 14:265–294, 1978.

[24] M. E. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(016128), 2002. DOI: 10.1103/PhysRevE.66.016128.

[25] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86(14):3200–3203, Apr 2001.

[26] Jonathan M. Read and Matt J. Keeling. Disease evolution on networks: the role of contact structure. *Proc. R. Soc. Lond. B*, 270:699–708, 2003.

[27] Everett M. Rogers. *Diffusion of Innovations*. Simon & Shuster, Inc., 5th edition, 2003.

[28] D. I. Rubenstein, S. Sundaresan, I. Fischhoff, and D. Saltz. Social networks in wild asses: Comparing patterns and processes among populations. In A. Stubbe, P. Kaczensky, R. Samjaa, K. Wesche, and M. Stubbe, editors, *Exploration into the Biological Resources of Mongolia*, volume 10. Martin-Luther-University Halle-Wittenberg, 2007. In press.

[29] Siva R. Sundaresan, Ilya R. Fischhoff, Jonathan Dushoff, and Daniel I. Rubenstein. Network metrics reveal differences in social organization between two fission-fusion species, Grevy's zebra and onager. *Oecologia*, 2006. doi 10.1007/s00442-006-0553-6.

[30] M. Tsvetovat, K. Sycara, Y. Chen, and J. Ying. Customer coalitions in electronic marketplaces. In Ulises Cortes Frank Dignum, editor, *Agent-Mediated Electronic Commerce III, Lecture Notes on Artificial Intelligence*. Springer-Verlag, 2003.

[31] Damián H. Zanette. Dynamics of rumor propagation on small-world networks. *Phys. Rev. E*, 65(4):041908, Mar 2002.