A Bi-Threshold Model of Complex Contagion and its Application to the Spread of Smoking Behavior

Chris Kuhlman, V. S. Anil Kumar, Madhav Marathe, Samarth Swarup, Gaurav Tuli
Network Dynamics and Simulation Science Lab, Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA 24060.
{ckuhlman, akumar, mmarathe, swarup, gtuli}@vbi.vt.edu

S. S. Ravi, Daniel J. Rosenkrantz
Department of Computer Science, University at Albany, State University of New York, Albany, NY 12222.
{ravi, djr}@cs.albany.edu

ABSTRACT

We study the dynamics of a bi-threshold model of contagion, wherein each node can be in one of two states (0 or 1), and will only change state if a minimum number (specified by an up-threshold and a down-threshold at each node) of its neighbors are in the opposite state. This model applies to processes where peer pressure is a strong factor in behavior change in either direction, such as initiation and cessation of smoking among adolescents.

We investigate this model both theoretically and experimentally. On the theoretical side, we establish results which show significant differences between simple contagions (where all thresholds are 1) and complex contagions (where one or more thresholds exceed 1) with respect to the complexity of determining several global properties of the system. On the experimental side, we apply this model to the data about adolescent smoking behavior from the National Longitudinal Study of Adolescent Health (Add Health) to analyze network dynamics such as the rate of spread and outbreak size.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms

Algorithms, Experimentation, Theory

Keywords

Complex contagion, bi-threshold model, adolescent smoking, synchronous dynamical systems

1. INTRODUCTION

Many social phenomena, such as smoking behavior, spread of information, diseases and viral marketing, can be modeled as contagion processes [6, 9, 12, 18, 22, 24, 34], in which an individual's state or choice is influenced by her neighbors' choices, leading to cascading effects. These models are predominantly **progressive** models [23], where in a two-state $\{0, 1\}$ system, nodes can only transition from 0 to 1; not 1 to 0. One way to capture this influence is through thresholds (e.g., [6, 22]). Informally, if a sufficient (i.e., a threshold) number of one's neighbors behave in a particular way, the node will also adopt that behavior.

Many real-world issues are characterized by two-choice back-and-forth decisions where a node can change state back and forth between 0 and 1. Schelling refers to this as cyclic behavior (at the micro and macro levels), and states: "Numerous social phenomena display cyclic behavior ..." (p. 86 [28]). He goes on to present everyday examples such as whether pick-up volleyball games early in the fall semester at Harvard will continue through the semester or die (e.g., individuals regularly choosing to play or not play). In refering to repeated decisions as to whether people will cross a street against traffic lights, a threshold is used explicitly in Schelling's description: "At some point [after some have walked into the street], several appear to decide that the flow of pedestrians is large enough to be safe and they join it, enlarging it further and making it safe for a few who were still waiting and who now join in." (p. 92 [28]). He also describes how people may initially step out into the street, but will retreat if there is an insufficient number of followers. One can also look at public health concerns such as obesity, where an individual's back-and-forth decisions to diet or not are so commonplace that it has a name: "yo-yo dieting" [1]. Moreover, dieting decisions are peer-influenced [7]. The point is that back-and-forth threshold systems are prevalent, and as will be described, these systems are also applicable to smoking.

The focus of our paper is studying a model for smoking behavior in adolescents through social network analysis and agent-based simulation (ABS). According to the World Health Organization (WHO) [33], smoking is responsible for 10% of all adult deaths and is the leading cause of preventable deaths. According to The Centers for Disease Control and Prevention, direct health care costs of smoking and produc-

tivity losses from adverse health are $96 billion and $97 billion annually, respectively [29].

A large number of factors are involved in the risk for smoking initiation among adolescents, including place of residence [2], schools and peer networks [8, 31], parental and familial influences [16], media messages [32], cigarette prices and policies [26], socioeconomic factors [5], and biological and cognitive factors [10]. However, Hoffman et al. recently reviewed the literature on adolescent cigarette smoking and found that peer influence, typically measured as the number of friends who smoke, has been repeatedly found to be the strongest risk factor [21]. Other ways of measuring peer influence, such as embeddedness in friendships, friendship quality, and peer social status also confirm this basic picture [13]. Recently, Go et al. [17] examined the Add Health data and showed that peer influence is a factor in both initiation and cessation of smoking among adolescents. Since peer influence is the biggest factor in the spread of smoking behavior, our model is especially appropriate for its study.

Christakis and Fowler showed through analysis of the Framingham Heart Study data that people tend to both start and stop smoking in groups [8]. Simple independent cascade models are unlikely to be valid for smoking behavior, because they cannot ensure such a property. Complex contagion models, in which a node switches from state 0 to state 1 if the number of neighbors in state 1 exceeds a threshold, can exhibit such a property. However, these models are monotone, and cannot explain other effects such as cessation of smoking.

In this paper, we propose a novel bi-threshold model of complex contagion, where transitions at each node are governed by two threshold values. When a node is in state 0, it transitions to state 1 when the number of its neighbors in state 1 equals or exceeds its *up-threshold*, denoted t_{up}. On the other hand, when a node is in state 1, it transitions to state 0 when the number of its neighbors in state 0 equals or exceeds its *down-threshold*, denoted by t_{down}. We apply the bi-threshold model of complex contagion to data from the National Longitudinal Study of Adolescent Health (Add Health [20]) on the spread of smoking behavior through adolescent friendship networks. Our main contributions are:

1. The bi-threshold model captures a number of effects observed in smoking behavior, including changes in groups and non-monotonic behavior, as we discuss later. The dynamics of such models are very complex and sensitive to the threshold values and underlying networks, which we verify empirically as well.

2. We study the complexity of the problems of computing dynamical properties of bi-threshold systems, e.g., reachability and fixed points, and show that they are NP-complete or #P-complete, in general. For systems with up and down threshold values of 1, we characterize the dynamics completely if the underlying contact graph is undirected. However, the problems become harder for directed graphs.

3. We infer parameters for a bi-threshold model to fit the available data on smoking status in waves I and II of the Add Health survey. This survey only has smoking states for a subset of nodes, which allow us to infer either an up or down threshold for them. We use a regression analysis to infer the initial states and the remaining thresholds for all the nodes.

4. We experimentally characterize the dynamical properties of such systems, and find that they generally converge to a set of pseudo-stationary configurations, in which the number of nodes in state 1 (i.e., the smoking state) does not vary much.

5. We find that high out-degree nodes have a significant impact on the rate of prevalence or cessation of smoking. If a small fraction (about 3.5%) of the highest out-degree nodes become smokers, the fraction of nodes in state 1 is almost doubled. On the other hand, if the same fraction of nodes becomes non-smokers, the fraction of nodes in state 1 reduces to a third. This corroborates with the observation that peer-influence is one of the most significant determinants of smoking behavior [8, 21].

Thus, the bi-threshold model generalizes a number of earlier threshold-based models for contagion. It captures complex non-monotonic diffusion phenomena, such as smoking behavior, much more realistically than other contagion models. Formal validation of the applicability of this model in the context of smoking behavior is difficult because of lack of adequate longitudinal smoking behavior data. This is a general problem with social phenomena that occur on a timescale of years, because gathering a large enough sample over many years is an effortful enterprise. For instance, the Add Health data only has partial information of the smoking states of a subset of the population studied over waves I and II. Nonetheless, our results with the networks in the Add Health data show many of the qualitative features observed by empirical studies on smoking behavior (which are not based on networked processes). Our approach is a first step towards providing a formal framework to unify and explain the diverse results on smoking behavior, and can help in evaluating and designing policies for controlling its spread. Further, smoking is a prototypical example, and such a framework would be useful in other applications.

Organization. In Section 2, we formally describe the bi-threshold model, following which we describe related work (Section 3) and then prove several theoretical results about the bi-threshold model in Section 4. In Section 5, we describe the Add Health data and our methodology for building the bi-threshold model to fit these data. We numerically explore the behavior of the bi-threshold model on the five largest friendship networks from the data. Comparing data from wave I and wave II allows us to construct a model for estimating threshold values for each node in the networks. In Section 6, we perform ABS with these estimated threshold parameters to study the dynamics of this model.

2. MODEL DESCRIPTION

2.1 Definition of the Bi-threshold Model

We model the propagation of contagions over a social network using discrete dynamical systems (e.g. [3]). We begin

with a definition of the model. Let $\mathbb{B}$ denote the Boolean domain {0,1}. A **Synchronous Dynamical System** (SyDS) $\mathcal{S}$ over $\mathbb{B}$ is specified as a pair $\mathcal{S} = (G, \mathcal{F})$, where (a) $G(V, E)$, an undirected graph with n nodes, represents the underlying social network over which the contagion propagates, and (b) $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$ is a collection of functions in the system, with f_i denoting the **local transition function** that computes the next state of v_i, $1 \leq i \leq n$.

Each function f_i specifies the local interaction between node v_i and its neighbors in G. (We use the convention that a node is *not* a neighbor of itself.) To provide additional details regarding these functions, we note that each node of G has a state value from $\mathbb{B}$. The inputs to function f_i are the state of v_i and those of the neighbors of v_i in G; function f_i maps each combination of inputs to a value in $\mathbb{B}$. In this paper, function f_i at node v_i is a **bi-threshold function**, characterized by two non-negative integer values denoted by $t_{\text{up}}(v_i)$ and $t_{\text{down}}(v_i)$. A precise definition of the function f_i is as follows.

(a) If the state of v_i is 0, then f_i is 1 if at least $t_{\text{up}}(v_i)$ of the neighbors of v_i are in state 1; otherwise, the value of f_i is 0.

(b) If the state of v_i is 1, then f_i is 0 if at least $t_{\text{down}}(v_i)$ of the neighbors of v_i are in state 0; otherwise, $f_i = 1$.

Thus, $t_{\text{up}}(v_i)$, called the **up-threshold** of v_i, represents the minimum number of neighbors of v_i that must be in state 1 for v_i to change from 0 to 1. Likewise, $t_{\text{down}}(v_i)$, called the **down-threshold** of v_i, represents the minimum number of neighbors of v_i that must be in state 0 for v_i to change from 1 to 0. A SyDS in which each node has a bi-threshold transition function is called a **bi-threshold** SyDS, denoted by BT-SyDS.

A **configuration** $\mathcal{C}$ of a SyDS at any time is an n-vector $(s_1, s_2, \ldots, s_n)$, where $s_i \in \mathbb{B}$ is the value of the state of node v_i. A single SyDS transition from one configuration to another can be expressed by the following pseudocode.

for each node v_i **do in parallel**
(i) Compute the value of f_i. Let s_i' denote this value.
(ii) Update the state of v_i to s_i'.
end for

Thus, in a SyDS, nodes update their state *synchronously*. Other update disciplines (e.g. sequential updates) for discrete dynamical systems have also been considered in the literature [3].

If a SyDS has a transition from configuration $\mathcal{C}_1$ to configuration $\mathcal{C}_2$, we say that $\mathcal{C}_2$ is the **successor** of $\mathcal{C}_1$ and that $\mathcal{C}_1$ is a **predecessor** of $\mathcal{C}_2$. A configuration that has no predecessor is called a **garden of eden** (GE) configuration. A configuration $\mathcal{C}$ is called a **fixed point** if the successor of $\mathcal{C}$ is $\mathcal{C}$ itself.

One can also define the bi-threshold model where the underlying graph of the dynamical system is *directed*. This is useful in modeling diffusion processes where the the influence relation between pairs of nodes is not symmetric; for example, a node u may influence another node v, but v may not have any influence on u. This asymmetry holds in the context of peer influence which is known to play a major role in the smoking behavior of adolescents (e.g., see [21]). If u influences v, the underlying graph contains the directed edge (u, v). The threshold values are with respect to the in-neighbors of a node v, that is, the set of nodes which have a directed edge to v. Thus, if a node v has an up-threshold $t_{\text{up}}(v)$, then at least $t_{\text{up}}(v)$ of its in-neighbors must be state 1 for v to change from 0 to 1. A similar explanation holds for the down-threshold values.

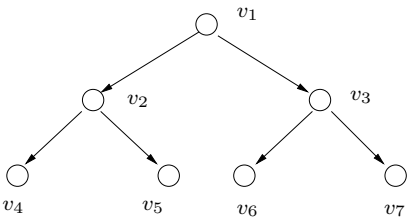


Figure 1: A directed BT-SyDS to illustrate an observed behavior.

Motivation: modeling smoking behavior. We now present a toy example to point out that the bi-threshold model with directed graphs can capture some observed behaviors in the context of smoking. In the literature, it is noted that "over time smokers were more likely to appear at the periphery" of the underlying social network [8]. To see how this behavior can occur under the bi-threshold model, consider the directed tree shown in Figure 1. Initially, the root node v_1 is in state 1 (corresponding to a smoker) and all the other nodes are in state 0 (corresponding to non-smokers). For each node, up-threshold is set to 1. For v_1, the down threshold is set to 0; for v_2 and v_3, the down threshold is set to 1, and for nodes v_4 through v_7, the down threshold is set to 2. It can be verified that during the first time step, v_1 changes to 0, v_2 and v_3 change to 1, while the other nodes remain at 0. During the second time step, v_1 remains at 0, v_2 and v_3 change to 0 and the other nodes change to 1. This is a *fixed point* of the system, and the nodes in state 1 correspond to leaves which are at the periphery of the graph. Also, observe that nodes v_2, v_3 simultaneously switch to state 1 in time step 2, and then switch to state 0 in time step 3. This captures the observation of Christakis and Fowler [8], that people tend to both start and stop smoking in groups.

2.2 Additional Definitions Related to the Model

For any SyDS $\mathcal{S}$, the **phase space** of $\mathcal{S}$ is a directed graph with one node for each possible configuration; there is a directed edge from the node representing configuration $\mathcal{C}$ to that representing configuration $\mathcal{C}'$ if and only if $\mathcal{C}'$ is the successor of $\mathcal{C}$. Since the domain is $\mathbb{B} = \{0,1\}$ and the underlying graph has n nodes, the number of nodes in the phase space is 2^n; thus, the size of the phase space is *exponential* in the size of the SyDS.

As defined above, the SyDS model is deterministic; that is, each configuration has a unique successor. Thus, the outdegree of each node in the phase space is 1. Each fixed point of a SyDS $\mathcal{S}$ is a self loop in the phase space of $\mathcal{S}$. Also, for any GE configuration, the corresponding node in the phase space has its indegree equal to zero.

A BT-SyDS in which $t_{\text{up}}(v) = t_{\text{down}}(v) = 1$ for each node v is called a **simple** BT-SyDS. If at least one of the threshold values in $\mathcal{S}$ is *greater than* 1, then $\mathcal{S}$ is referred to as a **complex** BT-SyDS.

If $t_{\text{up}}(v) = 0$ for some node v, then the state of v changes from 0 to 1 even when none of the neighbors of v is 1. We call such a node v an **uncontrolled up node**. Likewise, if $t_{\text{down}}(v) = 0$ for some node v, then node v will be referred to as an **uncontrolled down node**.

2.3 Computational Problems for BT-SyDSs

We study a number of different computational problems for BT-SyDSs. These problems model several questions regarding the global behavior of the underlying social network. Our results show a number of interesting differences between the complexities of these problems for simple and complex BT-SyDSs. Formal definitions of the problems studied in this paper are given below. In the literature, some of these problems have been considered for other dynamical system models (e.g. [3]).

Given a BT-SyDS $\mathcal{S}$, the **Fixed Point Existence** (FPE) problem asks whether $\mathcal{S}$ has a fixed point. The corresponding counting problem (i.e., finding the number of fixed points of $\mathcal{S}$) is denoted by #FPE.

Given a BT-SyDS $\mathcal{S}$ and a configuration $\mathcal{C}$, the **Predecessor Existence** (PRE) problem asks whether the configuration $\mathcal{C}$ has a predecessor. The corresponding counting problem (i.e., finding the number of predecessors) is denoted by #PRE.

Given a BT-SyDS $\mathcal{S}$ and two configurations $\mathcal{C}_1$ and $\mathcal{C}_2$, the **Configuration Reachability** (REACH) problem asks whether the system can reach $\mathcal{C}_2$ starting from $\mathcal{C}_1$.

Analytical results for the above problems are presented in Section 4.

3. RELATED WORK

Works on smoking behavior were presented in Section 1 and are not repeated. Here, we focus on contagion models and complexity results for contagion processes.

Different models have been used to study contagion processes, such as the independent cascade model, complex contagion [6] and its special case, the linear threshold model [22], and the linear influence model (LIM) [34]. Important questions in such applications include understanding steady state behavior, identifying influential individuals and the impact of network structure, and designing strategies to control the spread [12, 19, 22]. Yang et al., [34] use the linear influence model to fit the dynamics of information diffusion in Twitter. These models are so-called progressive models in which the only state transition allowed in a two-state $\{0, 1\}$ system is from state 0 to state 1.

Back-and-forth models also include the transition from state 1 to state 0. In the voter model [14], a node assumes the state of one randomly chosen neighbor, so that not all neighbors provide influence at each time, as does our model. In majority models (e.g., [11]), if one-half or more of a node's neighbors are in the opposite state, then a node transitions to that state. Our model is more general in that we can specify any minimum number of neighbors required to cause state transition, which may be more or less than one-half of nodes. We can also use relative thresholds, where a node's absolute threshold is normalized by its degree. A back-and-forth model is presented in [4]. However, that uniform mixing model assumes every node is connected to (and influenced by) all other nodes, so the graph forms a clique, and therefore that all information is globally known. In contrast, our system takes into account a population's connectivity, so that a node's interactions are local (confined to its neighborhood), which is more realistic in many cases. If global knowledge is required, we merely add a node and connect it to all others.

There are numerous results for computational complexity. In [22], the problem of finding the set of nodes of maximum size β that will result in the most nodes reached by a progressive diffusion process is shown to be NP-hard. A series of extensions is provided in papers up through [27]. See [25] for results on blocking diffusion. Results for the voter model and majority model are given in [14] and [11], respectively. We know of no complexity results for bithreshold models.

4. PHASE SPACE PROPERTIES OF BT-SyDSs: COMPLEXITY RESULTS

In this section, we present results on the complexity of testing various phase space properties for bi-threshold systems. The results are presented for the case of undirected graphs. By replacing each undirected edge $\{u, v\}$ by the pair of directed edges (u, v) and (v, u), it can be seen that the hardness results carry over to the directed case as well. We also discuss an interesting difference between the undirected and directed graph models with respect to phase space properties.

Due to space limitations, full proofs will appear in a complete version of this paper.

4.1 Fixed Point Existence and Counting

If a BT-SyDS does not have any uncontrolled up (down) nodes, then the configuration of all 0's (1's) is a fixed point. Therefore, the NP-hardness results given below for the FPE problem are for BT-SyDSs which contain nodes with up-threshold = 0 as well as those with down-threshold = 0.

THEOREM 4.1. *(i) The* FPE *and* #FPE *problems can be solved efficiently for BT-SyDSs with a maximum threshold of 1.*
(ii) The FPE *and* #FPE *problems are NP-complete and #P-complete respectively for BT-SyDSs where the maximum threshold is 2.*

Proof sketch: Part (i) is proven by considering each connected component (CC) of the underlying graph and showing that there are at most two candidate fixed points for the CC. Part (ii) is proven by a reduction from a restricted version of the Boolean Satisfiability problem (SAT), where each clause contains two or three literals and each literal appears in one or two clauses. (This restricted version of SAT is known be NP-complete [15].) ■

One may also consider variants of the FPE problem such as the following: Given a BT-SyDS $\mathcal{S}$ and an integer $q \leq n$, does $\mathcal{S}$ have a fixed point in which at most q nodes are in state 1? This variant, which we denoted by MIN-1-FPE, is motivated by the problem of determining whether a given social network has a stable configuration with only a small number of nodes in state 1. The following results hold for MIN-1-FPE.

THEOREM 4.2. *(i) The* MIN-1-FPE *problem can be solved efficiently for BT-SyDSs with a maximum threshold of 1.*
(ii) The MIN-1-FPE *problem is NP-complete for BT-SyDSs where the maximum threshold is 2.*

Proof sketch: Part (i) is proven by constructing a configuration $\mathcal{C}$ such that the only nodes which are in state 1 in $\mathcal{C}$ are those nodes which must be in state 1 in any fixed point of $\mathcal{S}$. The answer to the MIN-1-FPE instance is "yes" if and only if $\mathcal{C}$ is a fixed point and the number of nodes in state 1 in $\mathcal{C}$ is at most q. Part (ii) follows immediately from the NP-completeness of the FPE problems for BT-SyDSs in which the maximum threshold value is 2 (Part (ii) of Theorem 4.1) by setting $q = n$. ■

The above theorems provide a clear delineation between the complexities of FPE and #FPE problems for simple and complex contagions.

4.2 Predecessor Existence and Counting

THEOREM 4.3. *(i) The* PRE *problem can be solved efficiently for BT-SyDSs with a maximum threshold of 1.*
(ii) The #PRE *problem is #P-complete even when the maximum threshold is 1.*
(iii) The PRE *problem is NP-complete for BT-SyDSs even when all thresholds are 2.*

Proof (idea): Part (i) is proven by a careful analysis of the phase space of the system. (The proof involves several lemmas that capture how thresholds and state values of neighbors of a node v determine the state value of v in the predecessor configuration, if such a configuration exists.)

Part (ii) is shown by a parsimonious reduction from the problem of counting the number of satisfying assignments to 2CNF formulas which contain only positive literals. (The latter problem was shown to be #P-complete in [30].)

Part (iii) is shown by a reduction from a special version of SAT mentioned in the proof sketch for Theorem 4.1. ■

4.3 Reachability Problem

THEOREM 4.4. *The* REACH *problem can be solved efficiently for BT-SyDSs with a maximum threshold of 1.*

Proof (idea): This theorem is also proven by a careful analysis of the phase space of the system. In particular, the proof shows that BT-SyDSs in which the maximum threshold value is 1 either reach a fixed point or a 2-cycle in the phase space after a number of transitions that is bounded by the diameter of the underlying graph. ■

Like the FPE problem, one may also consider variants of the REACH problem. One such variant is the following: Given a BT-SyDS $\mathcal{S}$ an initial configuration $\mathcal{C}$ and an integer $q \leq n$, does the system reach a configuration in which at most q nodes are in state 1? If nodes in state 1 correspond to smokers, this question asks whether a social network will reach a configuration with a small number of smokers. This variant of reachability can also be solved efficiently for BT-SyDSs in which the maximum threshold value is 1 using the proof idea for Theorem 4.4 mentioned above.

Whether the REACH problem and its variant mentioned above can be solved efficiently for complex BT-SyDSs remains an open question. Below, we investigate it through simulation.

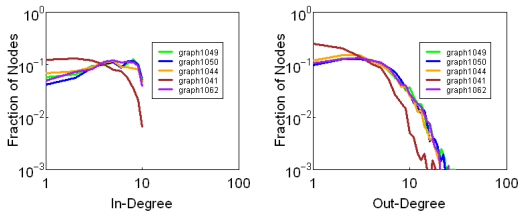
4.4 Undirected and Directed Graph Models

We now point out an interesting difference between the phase spaces of BT-SyDSs under undirected and directed graph models. As mentioned in the proof sketch for Theorem 4.4, for undirected graphs, when each threshold value is at most 1, every directed cycle in the phase space has length at most 2. This property doesn't hold when the underlying graph is directed. To see this, consider a BT-SyDS where the underlying graph is a simple directed cycle with $n \geq 3$ nodes. Let $V = \{v_1, v_2, \ldots, v_n\}$ denote the nodes and let $A = \{(v_1, v_2), (v_2, v_3), \ldots, (v_{n-1}, v_n), (v_n, v_1)\}$ be the edge set. Assume that the up and down threshold for each node is 1. In the initial configuration, let the state of v_1 be 1 and the states of the other nodes be 0. It can be verified that during successive time steps, the 1 value circulates among the nodes of the system, creating a cycle of length n in the phase space. One can construct other examples of BT-SyDSs with directed graphs such that their phase spaces have even longer cycles. Such examples point out that there are significant differences between the phase spaces of BT-SyDSs with undirected and directed graphs.

We now apply the bi-threshold model to smoking data from Add Health.

5. PARAMETER ESTIMATION

We focus on waves I and II of the Add Health survey [20]. Wave I data, collected in 1994-95, contain 85 adolescent friendship networks which were obtained by asking students at participating middle and high schools to nominate their friends. The networks contain 72589 distinct IDs. Of these, in-home interviews were conducted with 20746 individuals. For the present study, the variables we looked at were: whether the individual had smoked in the last 30 days, whether either of the individual's parents smoke, and the individual's age and gender. In wave II, follow-up interviews were conducted (in 1995-96) with 14739 of these individuals, and the same data were gathered again.



(a) (b)

Figure 2: Frequency distributions of (a) indegree and (b) outdegree for five of the largest networks from the Add Health data set.

For the construction of the friendship networks, each person was asked to name up to five male and five female friends. We take these friends as influencers of the person; i.e., these friends influence the smoking behavior of the person. A directed edge (u, v) from node u to node v means that u influences v. Thus, a node's in-degree is the number of influencers for that node, and has a maximum value of 10. The maximum out-degree of any node was 36; i.e., some individuals influence 36 others. In-degree and out-degree distributions for the five largest networks, with numbers of nodes ranging from 2000 to 2600, are shown in Figure 2.

Since the variables of interest are not available for the individuals who were not interviewed at home, we first take some simple steps to fill in the missing data. From the $\sim$ 14000 nodes for which age is known, a histogram was generated. Ages were assigned to unknown nodes randomly from the resulting distribution. For nodes with unknown gender, male or female was assigned with 50% probability. Parents' smoking status (hereafter PS) is known for $\sim$17000 nodes; of these, 75% of nodes had PS value of 1 (i.e., at least one parent smokes). Nodes with unknown PS were assigned 1 with a 0.75 probability. This procedure was applied to the entire data set, rather than to data for individual networks, because for some networks the available data were meager.

We then performed a set of Poisson regressions, grouping nodes by age, and then using gender and PS to determine the probability of being a smoker. This was used to fill in initial (wave I) smoking states for nodes for which this variable was unknown. For each node, the initial state is decided based on $v = Poisson(\lambda)$, where $Poisson(\lambda)$ is the realization of a Poisson random number. If $v = 0$, the initial state is 0 (i.e., non-smoking) and if $v > 0$, the initial state is 1. Values for these fits for each age are given in Table 1. There the columns are the gender of the child and whether the parents smoke or not. This process of assigning age, gender, PS, and initial smoking state provides one collection of traits and initial state. To assess variability, this process was repeated 500 times. Hence, nodes with unknown traits or initial state were assigned different values across the 500 instances, but known values for nodes were always used. The initial smoking states are used directly to specify the initial configuration for each of 500 diffusion instances of simulations in the next section. Roughly two-thirds of nodes are initially in the non-smoking (0) state for each of the 500 instances. Both traits and initial state were also used to derive regressions for t_{up} and t_{down}, as described next. One threshold can be inferred for every node for which the

Age	F, PS = 0	M, PS = 0	F, PS = 1	M, PS = 1
13	0.08	0.08	0.2	0.2
14	0.12	0.14	0.3	0.33
15	0.23	0.23	0.47	0.47
16	0.24	0.25	0.47	0.5
17	0.27	0.32	0.53	0.63
18	0.31	0.39	0.52	0.65

Table 1: λ values by age, gender, and parents smoking status (PS) for deciding the initial smoking state of nodes in the networks. Values have been rounded to two places after the decimal point for presentation here.

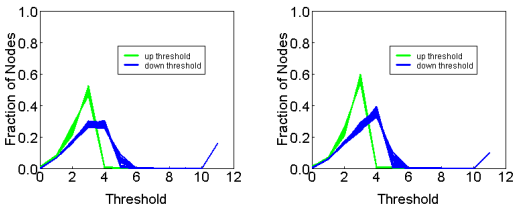


Figure 3: Twenty randomly chosen sets of up and down threshold distributions (out of the 500 total sets) for (a) network 1044 and (b) network 1049. Thresholds of 11 mean essentially an infinite threshold because the maximum indegree of any node is 10; such nodes are "pure influencers."

smoking states in wave I and wave II are known. The rules for determining thresholds are as follows. If a node is in state 0 in wave I and state 1 in wave II, then an up-transition has taken place. Hence, the up-threshold is at most the number γ of influencers in state 1 in wave I. The up-threshold is assigned uniformly at random from the interval $[0, \gamma]$. For nodes that are in state 1 in wave I and state 0 in wave II, the down-threshold is determined analogously. For nodes that are in state 0 in both wave I and wave II, the number of wave I influencers in state 1 is not sufficient to cause an up-transition, and hence the up-threshold for the node is at least $(\gamma + 1)$, and we use this value. Again, a similar argument is made for the down-threshold. Approximately 8100 thresholds were inferred in this manner.

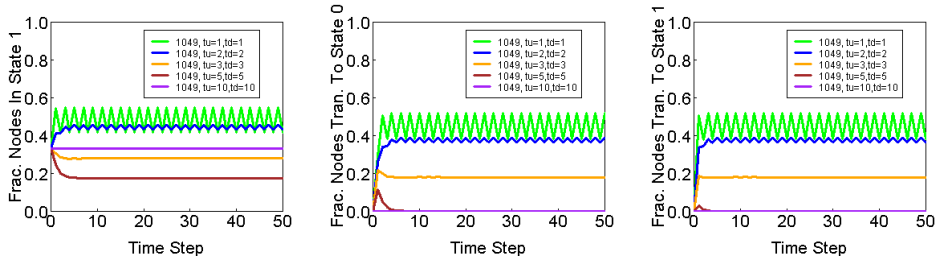
We now have for all nodes the values of the three traits and the initial smoking state, and for 8100 nodes, one threshold value. Linear regressions were performed to determine up-threshold as a function of the three traits and in-degree and out-degree for each of the 500 collections, and similarly for down-threshold, and these procedures yielded low variances. Up and down threshold distributions for 20 randomly chosen threshold assignments for two networks are given in Fig. 3.

6. SIMULATIONS AND RESULTS

We perform simulations to study multiple questions about the model. We use the largest five of the 85 networks from Add Health and the simulation parameters described in Section 5. We present results that elucidate the long-term dynamics of our model, and its sensitivity to different parameters. Then we conduct experiments where we freeze the highest out-degree nodes to study the impact of the thresholds of the hubs on the prevalence and cessation of smoking behavior in the network as a whole. We tie results to smoking behavior. Our main results are summarized below.

1. In our simulations we find that the system seems to converge to a "pseudo-stationary" state where the number of nodes in state 1 does not vary very much.

2. We find the results are very sensitive to the choices of thresholds. Increasing t_{up} and t_{down} uniformly (i.e., $t_{up} = t_{down} = t$) has interesting non-monotone effects. The fraction of nodes in state 1 decreases initially as t increases from 1 to 2, and then surprisingly *falls below* the fraction of 1-nodes in the initial configuration as t is further increased to 5. Note that this is



(a) (b) (c)

Figure 4: Dynamics for network 1049 for different $t_{up} = t_{down}$. (a) Fraction of nodes currently in state 1 as a function of time; (b) fraction of nodes transitioning to state 0; and (c) fraction of nodes transitioning to state 1. Figure (a) shows that the steady-state fraction of nodes in state 1 does not decrease monotonically to the initial fraction (at time 0), but rather as threshold increases to 3 and 5, the fractions of nodes in state 1 becomes less than the initial fraction.

counter to the behavior that would be observed if only the up-thresholds were increased, and thus the down-thresholds have an important role in this behavior.

3. We find that there can be large differences in the dynamics across different iterations, even if threshold values and proportion of initial smokers are held constant across iterations. The differences arise solely due to differences in initial conditions. These changes show up as subtle differences in average-case behavior, but show up as large differences when we plot the dynamics for all the individual iterations.

4. We find that nodes of high out-degree are highly influential in both prevalence and cessation of smoking rates. If a small fraction (3.5%) of the nodes with the highest out-degrees are permanent smokers (i.e., have states fixed at 1), the number of nodes in state 1 more than doubles. On the other hand, if this same set of nodes remain non-smokers (i.e., their states are fixed at 0), the fraction of nodes in state 1 becomes less than a third. This corroborates the observations that peer pressure has a significant impact on smoking [8, 21].

An **iteration** is a diffusion instance, where all nodes are assigned an initial state (either 0 or 1) and constant values for t_{up} and t_{down}. We only consider the first 50 time steps in the temporal evolution, where each step corresponds to one year, consistent with the one-year duration between wave I and wave II data sets in Add Health.

A **simulation** consists of a set of 100 or 500 iterations, each using a different set of initial state assignments. All the figures showing simulation results show curves which are calculated as point-wise averages from the set of iterations. Some of the figures (as will be noted) also show the dynamics of the individual iterations. Nodes whose initial states are known have the same initial state in all iterations; the remaining nodes are assigned a state according to their traits (age, gender, and whether their parents smoke), which can vary across iterations, as described in Sec. 5. Similarly, nodes for which a threshold can be inferred had the same inferred threshold over all iterations, but the other threshold for such nodes (either t_{up} or t_{down}), as well as both thresholds for other nodes, were assigned based on regressions that incorporate node traits. We discuss these experiments below.

1. Sensitivity to up and down threshold values. Nodes of networks were assigned homogeneous thresholds with $t_{up} = t_{down} = 1, 2, 3, 5$, and 10 for five simulations. Initial states were taken from the data sets. Average results in time over 100 iterations for network 1049 are provided in Figure 4. There is an initial transient phase, followed by a quasi-steady state phase where the behavior is approximately periodic. One would intuitively expect that as thresholds increase, the numbers of state transitions would decrease, and hence that the number of nodes in state 1 would monotonically approach that of the initial state configuration for all times. Figure 4(a) clearly shows that this is not the case. As threshold increases from 1 to 2, the numbers of nodes in state 1 decreases, but when thresholds increase to 3 and 5, the numbers of nodes in state 1 decrease below that of the initial configuration, before returning to the numbers of nodes in the initial configuration for threshold 10. Figures 4(b) and 4(c) show that the relative numbers of nodes transitioning to state 0 and state 1, respectively, are different for different thresholds in the transient regime, and these differences dictate whether the steady state fraction of nodes in state 1 is more or less than that of the initial configuration. These results are representative of the networks studied.

Note that the oscillations for $t_{up} = t_{down} = 1$ represent extreme behavior, perhaps applicable only to situations like street crossings in Section 1. However, it is interesting how increasing the thresholds by one attenuates the oscillations.

2. Variation across networks. Applying the heterogeneous thresholds and initial states of nodes as presented in Section 5 yields differences in behaviors across networks, exemplified by the results of Figure 5, which are average curves over 500 iterations. Network 1044 shows a decrease in numbers of nodes in state 1 over time, while other networks show an initial decrease, followed by an increase. The other two plots show that the numbers of nodes transitioning to state 0 are greater at the beginning of a simulation compared to the numbers changing to state 1, but that over time, the state transitions to 1 gradually increase. The plots also indicate that the numbers of nodes transitioning can fluctuate over 10 to 20 or more time steps (years) before settling to approximately steady state values. These swings are: (i) qualitatively similar to gradual changes observed in smoking, and (ii) illustrate that the model can produce "overshooting," which is important for social applications [4, 28].

3. Variance in long-term dynamics. Thus far, average behaviors have been emphasized to illustrate high-level trends.

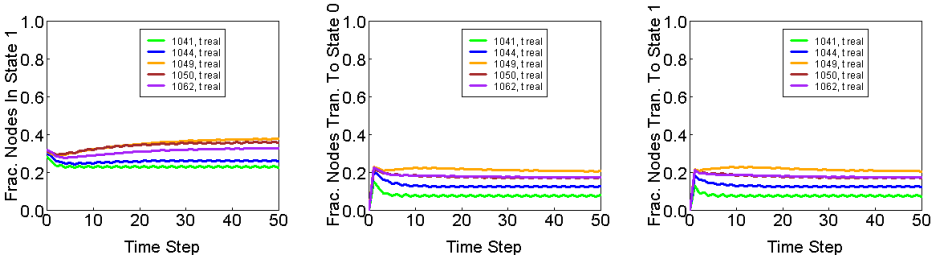


Figure 5: Dynamics for five networks for heterogeneous threshold values mined and inferred from observations. (a) Fraction of nodes currently in state 1; (b) fraction of nodes transitioning to state 0; and (c) fraction of nodes transitioning to state 1. These figures show that the average fractions of nodes in state 1 can either increase or decrease, or both, at early times, with transients out beyond time 20.

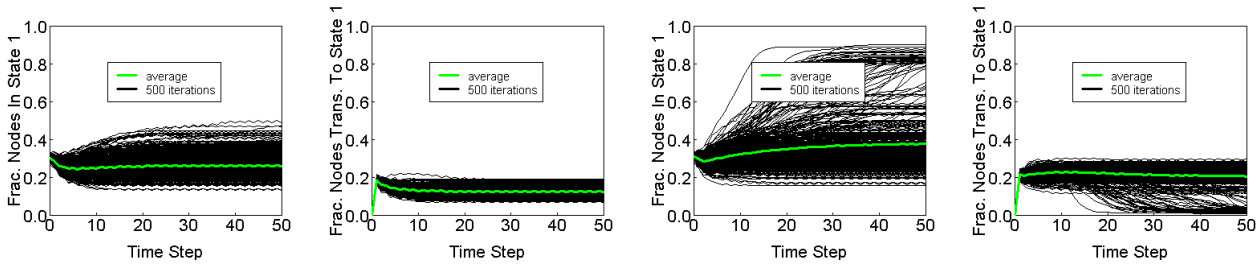
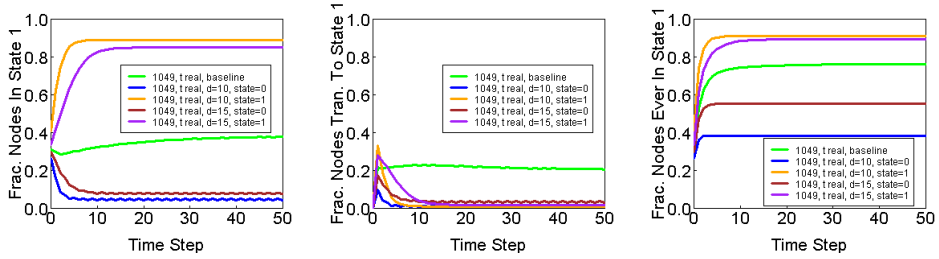


Figure 6: Dynamics for two networks showing 500 individual diffusion instances. (a) Fraction of nodes currently in state 1 for network 1044; (b) fraction of nodes transitioning to state 1 for network 1044; (c) fraction of nodes currently in state 1 for network 1049; (d) fraction of nodes transitioning to state 1 for network 1049. These results illustrate that focusing solely on average behavior hides ranges in behavior provided by individual diffusion instances.

However, average behaviors can hide variations in results across individual diffusion instances, and examples of this are depicted in Figure 6. The number of nodes in state 1 and the number of nodes transitioning to state 1 are provided for network 1044 in the first two plots, followed by corresponding plots for network 1049. Figures 6(a) and 6(c) show that differences in the average curves (in green) are driven primarily by diffusion instances for network 1049 that achieve 80% or more of nodes in state 1, and by diffusion instances for network 1044 that generate lesser fractions of nodes in state 1. Note that many instances between the two networks are within the same band. Figure 6(c) also indicates, through iterations whose curves are trending upwards at times between 40 to 50 steps, that transient behavior exists out to 50 years and beyond. Differences are also observed between Figures 6(b) and 6(d). The iterations in Figure 6(d) that tend to zero correspond to the iterations whose fraction of nodes in state 1 are 0.80 and above. There are no more nodes that can reach state 1, driving the number transitioning to zero. The plots for fractions of nodes transitioning to state 0 are very similar to those of Figures 6(b) and 6(d), so that the greater fractions in Figure 6(d) are offset by down transitions. Because the actual effects of smoking prevention policies correspond to one diffusion instance, policy makers must be aware that predicted average trends can hide complicated behaviors across iterations.

4. Relative influence of nodes in smoking prevalence and cessation. Correlation between the behavior of the popular students and of the group as a whole is well accepted, however causality can be argued in both directions. For example, Valente et al. [31] investigated the possibility that popular students are more susceptible to taking up smoking, by surveying middle-schoolers in southern California about their commitment not to smoke in the future. They found that popular students were statistically less likely to make the commitment. On the other hand, it can be argued that smoking becomes widespread in a network precisely if the popular students start smoking.

We present some simulations that show that the bi-threshold model exhibits a similar relationship between the states of the highest out-degree nodes and the fraction of smokers in the network. (The following results also control for the atmost q nodes in state 1, which is related to the variant of the REACH problem for complex BT-SyDSs in Section 4.3.) Specifically, we compare the effects of choosing a subset of nodes S according to two criteria: random and high out-degree based (i.e., consisting of the nodes with out-degree at least d, for different choices of d). In Figure 7, we show the effect of "freezing" the states of all nodes in S at value 1 (by setting their t_{down} to be very high values) for one of the networks we consider (refer to the curves labeled "state=1" in this figure); only the results for the choice of $d = 15$ and $d = 10$, corresponding to S having about 3.5% and 14.5% of the nodes, respectively, are shown here. We observe that freezing these top 3.5% high out-degree nodes results in a lot more nodes staying in state 1 (Figure 7(a)), compared to the baseline (which corresponds to the setting without any frozen nodes), and to random choice of a similar fraction of



(a) (b) (c)

Figure 7: For network 1049, (a) number of nodes currently in state 1, (b) number of nodes transitioning to state 1, and (c) cumulative number of nodes ever to reach state 1. The green curve is the baseline condition. The blue and brown curves fix the nodes that have out-degree at least 10 and 15, respectively, in state 0 and hence decrease the numbers of nodes in state 1 in the steady state and lower the fractions of nodes to ever reach state 1. The orange and purple curves, corresponding to fixing the same nodes in state 1, serve to increase the nodes in state 1. In fact, 3.5% of nodes fixed in state 1 (purple curve) is almost as effective as 14% of nodes fixed in state 1 (orange curve).

nodes (not shown here because of space constraints).

Similarly, the number of nodes which were ever in state 1 (which models individuals who might have ever smoked) is also much larger than the baseline (Figure 7(c)) and random. This suggests that if high out-degree individuals start smoking, it causes high prevalence of smoking. This raises a natural question: does this behavior also hold for smoking cessation? Indeed it does, as shown by the curves labeled "state=0" in Figure 7. Here, we freeze the states of the nodes in S to 0. Figure 7(a) shows that the number of nodes in state 1 at any time is much smaller than the baseline; further, the number of nodes ever in state 1 is also much smaller. As in the earlier scenario, we find that freezing 3.5% of the nodes to 0 is as effective as freezing 14.5% of the nodes at 0, with respect to the number of nodes in state 1 at any time. As mentioned earlier, the results for random choice make very minimal change in both scenarios, and are not shown here. These results are consistent with experimentally determined peer effects of smoking [8].

7. CONCLUSIONS

We present a novel model of complex contagion, the bi-threshold model, motivated by non-monotonic diffusion phenomena such as the spread of smoking behavior. Through theoretical and simulation results on networks from the Add Health survey, we find that this model captures a number of features of observed smoking behavior, such as the impact of peer-influence on smoking and its cessation. We also examine the computational complexity of determining fundamental dynamical properties of this model. The hardness results for absolute thresholds also hold for relative thresholds where thresholds are normalized by a node's in-degree; simulations can be run with relative thresholds as well. There are several directions for further research. Theoretically, it would be of interest to establish the complexity of the reachability problem for complex contagions, and to identify further differences in the structures of the phase spaces between undirected and directed graphs. From a practical perspective, a useful research direction is to identify and explore other contexts where the bi-threshold model (or a suitable generalization thereof) can be utilized.

Acknowledgments: We thank the reviewers, our external collaborators, and members of the Network Dynamics and Simulation Science Laboratory (NDSSL) for their suggestions and comments. We also acknowledge use of the Add Health data, administered by the University of North Carolina at Chapel Hill and funded by the Eunice Kennedy Shriver National Institute of Child Health and Human Development, among other agencies. We are especially indebted to Richard Beckman for his guidance and for developing the model for assigning initial smoking states. This work has been partially supported by NIH MIDAS project 2U01GM070694-7, NSF PetaApps Grant OCI-0904844, DTRA R&D Grant HDTRA1-0901-0017, DTRA CNIMS Grant HDTRA1-07-C- 0113, NSF Netse CNS-1011769, and NSF SDCI OCI-1032677.

8. REFERENCES

[1] R. Atkinson, W. Dietz, J. Foreyt, N. Goodwin, J. Hill, J. Hirsch, F. Pi-Sunyer, R. Weinsier, R. Wing, J. Hoofnagle, J. Everhart, V. Hubbard, and S. Yanovski. Weight Cycling. *Journal of the American Medical Association*, 272(15):1196–1202, 1994.

[2] J. R. Barnett. Does place of residence matter? Contextual effects and smoking in Christchurch. *The New Zealand Medical Journal*, 113(1120):433–435, 2000.

[3] C. L. Barrett, H. B. Hunt III, M. V. Marathe, S. S. Ravi, D. J. Rosenkrantz, and R. E. Stearns. Complexity of Reachability Problems for Finite Discrete Dynamical Systems. *J. Comput. Syst. Sci.*, 72(8):1317–1345, 2006.

[4] G. Bischi and U. Merlone. Global Dynamics in Binary Choice Models with Social Influence. *J. Math. Sociology*, 33:277–302, 2009.

[5] U. Broms, K. Silventoinen, E. Lahelma, M. Koskenvou, and J. Kaprio. Smoking cessation by socioeconomic status and marital status: The contributions of smoking behavior and family background. *Nicotine and Tobacco Research*, 6(3):447–455, 2004.

[6] D. Centola and M. Macy. Complex Contagions and the Weakness of Long Ties. *American J. Sociology*, 113(3):702–734, 2007.

[7] N. Christakis and J. Fowler. The Spread of Obesity in a Large Social Network Over 32 Years. *N. Engl. J.*

Med., pages 370–379, 2007.

[8] N. A. Christakis and J. H. Fowler. The collective dynamics of smoking in a large social network. *N. Engl. J. Med.*, 358(21):2249–2258, May 22 2008.

[9] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of KDD*, Las Vegas, Nevada, USA, August 24-27 2008.

[10] I. T. Croghan, J. O. Ebbert, R. D. Hurt, J. T. Hays, L. C. Dale, N. Warner, and D. R. Schroeder. Gender differences among smokers receiving interventions for tobacco dependence in a medical setting. *Addictive Behaviors*, 34(1):61–67, Jan 2009.

[11] P. Dreyer and F. Roberts. Irreversible k-Threshold Processes: Graph-Theoretical Threshold Models of the Spread of Disease and Opinion. *Discr. Appl. Math.*, 157:1615–1627, 2009.

[12] D. Easley and J. Kleinberg. *Networks, Crowds and Markets: Reasoning About A Highly Connected World*. Cambridge University Press, New York, NY, 2010.

[13] S. T. Ennett, R. Faris, J. Hipp, V. A. Foshee, K. E. Bauman, A. Hussong, and L. Cai. Peer smoking, other peer attributes, and adolescent cigarette smoking: A social network analysis. *Prevention Science*, 9:88–98, 2008.

[14] E. Even-Dar and A. Shapira. A Note on Maximizing the Spread of Influence in Social Networks. In *WINE 2007, LNCS 4858*, pages 281–286, 2007.

[15] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. W. H. Freeman and Co., San Francisco, CA, 1979.

[16] S. E. Gilman, R. Rende, J. Boergers, D. B. Abrams, S. L. Buka, M. A. Clark, S. M. Colby, B. Hitsman, A. N. Kazura, L. P. Lipsitt, E. E. Lloyd-Richardson, M. L. Rogers, C. A. Stanton, L. R. Stroud, and R. S. Niaura. Parental smoking and adolescent smoking initiation: An intergenerational perspective on tobacco control. *Pediatrics*, 123:e274–e281, 2009.

[17] M.-H. Go, H. D. Green Jr., D. P. Kennedy, M. Pollard, and J. S. Tucker. Peer influence and selection effects on adolescent smoking. *Drug and Alcohol Dependence*, 109:239–242, 2010.

[18] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2010.

[19] Habiba, Y. Yu, T. Berger-Wolf, and J. Saia. Finding Spread Blockers in Dynamic Networks. In *Proc. SNA-KDD Workshop*, 2008.

[20] K. Harris. The National Longitudinal Study of Adolescent Health (Add Health), Waves I and II, 1994-1996; Wave III, 2001-2002 [machine-readable data file and documentation], 2008. Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill 2008.

[21] B. R. Hoffman, S. Sussman, J. B. Unger, and T. W. Valente. Peer influences on adolescent cigarette smoking: A theoretical review of the literature. *Substance Use and Misuse*, 41:103–155, 2006.

[22] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the Spread of Influence Through a Social Network. In *Proc. ACM KDD*, pages 137–146, 2003.

[23] J. Kleinberg. Cascading Behavior in Networks: Algorithmic and Economic Issues. In *Algorithmic Game Theory*, chapter 24, pages 613–632. Cambridge University Press, NY, NY, 2007.

[24] G. Kossinets, J. Kleinberg, and D. Watts. The Structure of Information Pathways in a Social Communication Network. In *Proc. ACM KDD*, 2008.

[25] C. Kuhlman, V. Kumar, M. Marathe, S. Ravi, and D. Rosenkrantz. Finding Critical Nodes for Inhibiting Diffusion of Complex Contagions in Social Networks. In *Proc. ECML PKDD*, pages 111–127, 2010.

[26] L. Liang, F. Chaloupka, M. Nichter, and R. Clayton. Prices, policies and youth smoking, may 2001. *Addiction*, 98(Suppl 1):105–122, 2003.

[27] E. Mossel and S. Roch. On the Submodularity of Influence in Social Networks. In *Proc. ACM STOC*, pages 128–134, 2007.

[28] T. Schelling. *Micromotives and Macrobehavior*. W. W. Norton and Company, 1978.

[29] USA Today Newspaper. Do smokers cost society money?, 2009. 8 April 2009, `http://www.usatoday.com/news/health/2009-04-08-fda-tobacco-costs\_N.htm`.

[30] S. Vadhan. The Complexity of Counting in Sparse, Regular and Planar Graphs. *SIAM J. Comput.*, 31(2):398–427, 2001.

[31] T. W. Valente, J. B. Unger, and C. A. Johnson. Do popular students smoke? The association between popularity and smoking among middle school students. *Journal of Adolescent Health*, 37:323–329, 2005.

[32] M. Wakefield, B. Flay, M. Nichter, and G. Giovino. Role of media in influencing trajectories of youth smoking. *Addiction*, 98(Suppl 1):79–103, 2003.

[33] World Health Organization. Tobacco free initiative, 2010. `http://www.who.int/tobacco/health_priority`.

[34] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *Proceedings of the 10th IEEE ICDM*, Sydney, Australia, 2010.