General Threshold Model for Social Cascades: Analysis and Simulations

Jie Gao, Golnaz Ghasemiesfeh, Stony Brook University
Grant Schoenebeck, Fang-Yi Yu, University of Michigan

Social behaviors and choices spread through interactions and may lead to a cascading behavior. Understanding how such social cascades spread in a network is crucial for many applications ranging from viral marketing to political campaigns. The behavior of cascade depends crucially on the model of cascade or social influence and the topological structure of the social network.

In this paper we study the *general threshold model* of cascades which are parameterized by a distribution over the natural numbers, in which the collective influence from infected neighbors, once beyond the threshold of an individual u, will trigger the infection of u. By varying the choice of the distribution, the general threshold model can model cascades with and without the submodular property. In fact, the general threshold model captures many previously studied cascade models as special cases, including the independent cascade model, the linear threshold model, and k-complex contagions.

We provide both analytical and experimental results for how cascades from a general threshold model spread in a general growing network model, which contains preferential attachment models as special cases. We show that if we choose the initial seeds as the early arriving nodes, the contagion can spread to a good fraction of the network and this fraction crucially depends on the fixed points of a function derived only from the specified distribution. We also show, using a coauthorship network derived from DBLP databases and the Stanford web network, that our theoretical results can be used to predict the infection rate up to a decent degree of accuracy, while the configuration model does the job poorly.

Additional Key Words and Phrases: Social Cascades, General Threshold Model, Stochastic Attachment Graph, Preferential Attachment Graph

1. INTRODUCTION

Human activity is embedded in a network of social interactions, which can spread information, beliefs, diseases, technologies, and behaviors. A better understanding of these social interactions promises a better understanding of and the ability to influence a wide range of phenomena – financial practices [Banerjee et al. 2013; Coleman et al. 1957], healthy/unhealthy habits [Mermelstein et al. 1986], and voting practices [Adamic and Glance 2005], to name a few. In this paper we focus on social cascades, that start from a few nodes that are initially active or infected and spread through the edges of the network to other nodes. There are two important factors in determining the scope and rate of such diffusion: the model of contagions, i.e., how a node is influenced by its neighbors; and the network topology. We discuss these two factors separately.

Support: J. Gao and G. Ghasemiesfeh would like to acknowledge support through AFOSR FA9550-14-1-0193, NSF CNS-1217823, DMS-1418255, and CCF-1535900. G. Schoenebeck and F. Yu would like to acknowledge support through NSF CCF-1535912, CAREER CCF-1452915, a Facebook Faculty Award and a Google Faculty Award.

Author's addresses: J. Gao, G. Ghasemiesfeh, Computer Science Department, Stony Brook University; email: {jgao,gghasemiesfe}@cs.stonybrook.edu; G. Schoenebeck, F. Yu , Computer Science & Engineering Department, Univ. of Michigan; email: {schoeneb,fayu}@umich.edu.

EC'16, July 24–28, 2016, Maastricht, The Netherlands. ACM 978-1-4503-3936-0/16/07.

Social Contagion Models The study of contagions starts from the study of infectious diseases and epidemics [Jackson 2008]. Social behaviors and decisions are "contagious" too. The copying of behaviors leading to a social cascade of behavioral changes are attributed to two effects: the informational benefit (inferring hidden, private information others may know) and direct benefit effects (resulting from coordinated actions or social pressure).

The *general threshold model* [Granovetter 1978; Mossel and Roch 2007] is a fairly general model to capture such intuition. Each node v has a monotone function $g_v : \{0,1\}^{|\Gamma(v)|} \to [0,1]$, where $\Gamma(v)$ indicates the set of v's neighbors in a social network. The function g_v represents how much influence (via knowledge, social pressure, etc) any set of neighbors has upon node v. In the general threshold model, each node also has threshold th_v drawn uniformly and independently from the interval $[0,1]$. After an initial seed set is infected, a node v becomes infected if $g_v(S) \geq th_v$ where S is the set of infected neighbors of v.

The general threshold model captures many other models as special cases. For example, a special case is the *linear threshold model*, in which each edge (u,v) has an influence weight $w(u,v)$, and the function g_v is then the sum of the influence from all infected neighbors of v. [1] Another example of a class of cascades captured by the general threshold model is the independent cascade model [Goldenberg et al. 2001]. In this model, there is some fixed parameter ρ, and each infected node has one chance to infect each uninfected neighbor node with probability ρ (iid).

We call contagions *simple* when the influence g_v is submodular—that is $g_v(S' \cup \{x\}) - g_v(S') \leq g_v(S \cup \{x\}) - g_v(S)$, if $S \subseteq S'$—and call contagions *complex* when this fails to hold (e.g., contagions that require activation from multiple neighbors). In a simple contagion, the effect of an additionally infected neighbor is marginally decreasing. In a complex contagion, there could be an initial barrier such that no activation is possible until the barrier is crossed. There can be synergy between neighbors such that the total influence from them is not just a simple sum. If we define $f(S)$ as the expect number of infected nodes when the vertices in S are chosen as the initial seeds, then if g_v is submodular for all nodes, then f is submodular as well [Mossel and Roch 2010].

The monotonicity and submodularity have greatly helped with the analysis of the diffusion behavior with respect to the choice of seeds. In particular, one can apply the greedy set cover algorithm to choose the set of k best seeds to maximize the final scope of the contagion. This will give a $1 - 1/e$ approximation to the maximum scope obtained by any k seeds. In contrast, for the general threshold model, this is a very hard question and not much is known in the literature other than that is is NP-hard to even approximate [Kempe et al. 2003]. The two special cases, the linear threshold model and the independent cascade model, have received a lot of attention because they both have the submodular property [Kempe et al. 2003].

While this result has been well recognized and celebrated, a natural question one may ask is whether the submodularity assumption holds in reality and whether the result can be generalized. Sociologists observe that in the case of the adoption of pricey technology innovations, the change of social behaviors, and the decision to participate in a migration, etc [Coleman et al. 1966; Macdonald and Macdonald 1964], an additional confirmation is crucial, suggesting the model of complex contagion. In practice, threshold distributions are usually computed from data of contagions by using the empirical fraction of agents who adopt directly after k ties adopt, given that they had not previously. The distributions found depend on which cascades are analyzed, however,

[1] Often an additional restriction is imposed that for all nodes v: $\sum_{u \in \gamma(v)} w(u,v) \leq 1$ to ensure that g_v is always in $[0,1]$.

this conditional probability typically increases with k until some small constant of at least 2, and then then tapers off.Examples include LiveJournal [Backstrom et al. 2006], DBLP [Backstrom et al. 2006], Twitter [Romero et al. 2011], and Facebook [Ugander et al. 2012]. Some of these data sets indeed show diminishing return of the influence function, but others do not. They find that the second affected neighbor often has more marginal effect than the first. Additionally, the study in the Facebook data set shows that the number of *connected components* in the active neighbors is a much better predictor on the probability of joining Facebook, compared to the number of active neighbors.

Work done on complex contagions is much more limited and so far focused on a simplistic single threshold model called k-complex contagions. In k-complex contagions, all nodes have the same threshold k. A node becomes active if and only if at least k of its neighbors have been activated. It has been shown that a k-complex contagions is generally slower and more delicate than simple contagion $k = 1$ [Ebrahimi et al. 2014, 2015; Ghasemiesfeh et al. 2013]. One of the limitations of this k-complex contagion model is the dependency on the fixed threshold k for all nodes in the network. In practice there are people who like to try out new things and are more risk driven while others are risk averse. Therefore the threshold function is not necessarily uniform.

In this paper we consider one step of generalizing the k-complex contagion model by considering the threshold coming from a distribution D on positive intergers. The initial adoption barrier can still exist which makes the adoption function to be non-submodular. We provide analysis on the spreading behaviors on a general family of networks that grow over time.

Stochastic Attachment Network Model. In addition to a model of cascade, the model of network is also important. A lot of mathematical models have been developed to capture some of the attributes of real world social networks. A celebrated set of results are the family of small world graphs [D.Watts et al. 2002; D.Watts and S.Strogatz 1998; Kleinberg 2000, 2001; Newman and Watts 1999] and the family of graphs that produce power law degree distribution [Barabási and Albert 1999; Kleinberg et al. 1999; Kumar et al. 2000, 1999].

In this work we examine a growing network in which newcomers connect stochastically to nodes already in the network. This family of networks, which we call the *stochastic attachment network* model, has the preferential attachment network model as a special case. In the preferential attachment models [Barabási and Albert 1999], nodes arrive in a sequential order. Each node chooses m edges from the nodes that arrive earlier. When an edge is added, the neighbor is selected with probability proportional to its current degree. This model generates graphs with a power law degree distribution and has been used to explain the observation in web graphs and social networks. We examine a more general model in which new edges are not necessarily preferentially attached to existing nodes and each newcomer may have a varying number of edges. The key feature that is used in our analysis is that the network is formed over time, when new nodes arrive sequentially and attach to existing nodes.

We study contagions on both directed and undirected version of the stochastic attachment network. In the first case, we consider each edge issued by a newcomer u as directional, pointing to an earlier node v. This edge can be interpreted as u following edge v. A social contagion spreads in the *reverse* direction of an edge. This models information spreading in Twitter-type social networks, in which messages or information only travels along the direction of the edges. A node u will be influenced only by the neighbors u follows and not the neighbors that follow u. In the second case, all edges are treated as undirected, allowing contagions to spread in both directions. For example, consider a co-authorship network in which a new researcher choose to work with senior researchers/advisors, but here information or social influence is bidirectional.

An additional consideration is where the initially infected in the contagions reside within the network structure. In this paper we consider the scenario when some entity is trying to initiate a cascade. The entity is allowed to choose where the nodes go. We model this case by letting the seed equal the first nodes (in arrival order) or a subset of these nodes.

In our earlier work we show that due to the evolutionary nature k-complex contagions spread to the entire network in preferential attachment models and the contagion spreads very fast [Ebrahimi et al. 2014], when $k < m$ and the first few nodes in the arriving order are selected as the initial seeds in both the directed and undirected cases. This paper provides significant generalizations in both models of contagions and models of networks. The proof ideas are also completely new.

Our Results In this paper we study the behavior of a contagion following a general threshold model on both directed and undirected stochastic attachment graphs. We provide the most detailed analysis in the case of preferential attachment and later generalize to other scenarios.

We show that the number of infected nodes depends critically on the threshold distribution D. In the directed case, we derive a function $f : [0, 1] \to [0, 1]$ describing the probability of the i-th arriving node being infected, which depends only on a single number summarizing the status of the nodes with earlier arriving order, i.e., their threshold and whether they are infected or not. This function f has fixed points, which may be either stable or unstable. The ratio of the infected nodes in the network converges to one of these stable fixed points with high probability. When there are multiple fixed stable points, the contagion may converge to any one of them with at least constant probability.

In the undirected case, we note that the number of infected nodes will be no fewer than the directed case, since the edges can possibly spread social influence both ways. However, we show something much stronger than this, that, with high probability, the total number of infected nodes will always be a constant fraction higher than the highest stable fixed point of function f, when non-zero stable fixed points exist.

We performed both simulations and experiments with real world data sets. On various stochastic attachment graphs we observe the same behaviors as predicted in theory. We also tested real world networks. We used two datasets, the coauthorship derived from DBLP database which is an undirected graph and the Stanford web graph (which is naturally directed). On both datasets we infer the arriving order by using k-core decomposition – i.e., removing nodes with degree k for k starting from 1 recursively. We show that using the stochastic attachment model one can get fairly accurate prediction of the contagion rate. On the other hand, if we use the same degree distribution and generate a graph using the configuration model[2], on which the contagion behaviors differ significantly from that of the real netwrok. These experiments confirm the validity and utility of our model and analysis in helping to understand and predict contagions on real world graphs.

2. PRELIMINARIES

Definition 2.1. A ***General Threshold Contagion*** $\boldsymbol{GTC(G, D, I)}$ is a contagion which starts from a set of initial nodes I and spreads over the network G. Each node v has a threshold R_v which is drawn from distribution D for which the range is all positive integers. The contagion proceeds in rounds. At each round, each vertex v with at least R_v infected neighbors becomes infected.

[2] In a configuration model we fix the degree distribution first and then match the half edges at the nodes randomly.

Definition 2.2. The ***Stochastic Attachment Model***, $\mathrm{SA}_M(n)$ models a network with a growing number of vertices and edges. Denote by M the distribution of outgoing degree, with range between 1 and c_u and $E[M] = \mu_M$. We start with a complete graph on $c_u + 1$ nodes. At each subsequent time step t a node v arrives and adds m edges to the existing vertices in the network, where m is chosen from M. Denote the graph containing the first $n-1$ nodes as G_{n-1}. For each new vertex, we choose $w_1, w_2, \cdots, w_m$ vertices, possibly with repetitions from the existing vertices in the graph. Specifically, nodes $w_1, w_2, \cdots, w_m$ are chosen independently of each other conditioned on the past. For each i, w_i is selected from the set of vertices of G_{n-1} using an attachment rule $\mathbb{A}$. Then we draw edges between the new vertex and the w_i's. Repeated w_i's cause multiple edges. Also each node creates a self loop as well.

The stochastic attachment model is a general model that captures how nodes and edges are added over time. It contains the preferential attachment graph model as a special case, in which each node issues the same number of edges and a new edge attaches to a node u with probability proportional to u's current degree. In this case the degree distribution becomes power law. We remark that the stochastic attachment model is more general and may contain graphs with non-power law degree distribution (e.g., if the edges are attached uniformly at random).

In this paper we consider two cases, when the edges are considered directional or undirectional. In the directed case, each edge is issued by a node u and points towards a node v earlier in the arriving order. We consider this as u following v. Thus contagion propagates in the reversed direction of edge uv. A node u is infected if the number of infected nodes that u follows is greater than its threshold. In the undirected case, infection can happen in both directions. We denote the former as a directed contagion $GTD(G, D, I)$ on a graph G with initial seeds in I and threshold distribution D, and the latter as undirected contagion.

In this paper the initial seeds are chosen as a fraction (or all) of the first few nodes. When considered G as a graph in which the thresholds of the nodes are chosen when they appear, the contagion process is determined. In our analysis, however, we choose to delay revealing the thresholds of nodes in G. Due to space constraint, most of the technical proofs are put in the full version.

3. DIRECTED NETWORK

In this section, we analyze the number of nodes infected with a general threshold contagion model on a directed stochastic attachment graph. For the discussion below, we first focus on the case when each node chooses a fixed number of m edges for simplicity. At the end of this section we report how the results generalize to the other scenarios.

Since we are considering a directed contagion, we only have to consider the effect of outgoing edges of a node. There are at most m of them for each node u and these edges point to nodes that arrive earlier than u. Thus we can go through the list of nodes in their arriving order to calculate whether a node will be infected. Each node is only evaluated once. The first m nodes are the initial seeds I. We start at node of index $m + 1$ and process each of the following nodes in their order of arrival in the graph. When a node is being processed we reveal both its threshold and its outgoing edges, and based on its threshold and the status (being infected or not) of its outgoing edges, it is determined if the current node will be infected or not. To evaluate this probability we give some definitions.

Assume that node u is the i-th node in the arrival order in G. Let V_{i-1} be the set of first $i - 1$ nodes in G and X_{i-1} be the set of infected nodes in V_{i-1}. If u's threshold is $R_u = k$, u is infected if and only if among the m edges u issues, at least k of them land in nodes in X_{i-1}. Now consider a specific edge of u, we define Y_i as the probability that this edge lands in an infected node (e.g., in X_{i-1}). Y_i depends on the attachment rule $\mathbb{A}$ and the set of nodes that are infected so far. For example, if the edges of u are uniformly randomly selected among the nodes in V_{i-1}, then Y_i is the ratio of the infected nodes $|X_{i-1}|/|V_{i-1}|$. If the edges of u are preferentially attached, i.e., with probability proportional to the current degree of the nodes, Y_i is the ratio of the infected degree $Y_i = \sum_{v \in X_{i-1}} \deg(v)/\sum_{w \in V_{i-1}} \deg(w)$, Here $\deg(v)$ is the total degree of each node v (counting both incoming and outgoing edges).

Next we can compute the probability of node u being infected when its threshold is $R_u = k$. For that to happen, among the m edges of u, at least k of them need to land on a node in X_{i-1}. Now,

$$\mathrm{Prob}\{\text{Infection of } u | R_u = k\} = \sum_{\ell=k}^{m} \binom{m}{\ell} Y_i^{\ell}(1-Y_i)^{(m-\ell)} \tag{1}$$

Now, the probability of infection of node u is described by a function f:

$$\mathrm{Prob}\{\text{Infection of } u\} = f(Y_i) = \sum_{k} \mathrm{Prob}[R_u = k] \sum_{\ell=k}^{m} \binom{m}{\ell} Y_i^{\ell}(1-Y_i)^{(m-\ell)} \tag{2}$$

Therefore, the random process $\{Y_t : t = m+1, ..., n\}$ in $\mathrm{SA_M(n)}$, is a Markov chain that only depends on the previous state of the process. To understand the contagion we first need to understand this Markov process and in particular the function f. First f is a polynomial function. Thus it is continuous and differentiaable. It is also not hard to see that the function f is nondecreasing (with proof in the full version). Second, f maps values of $[0,1]$ to image domain $[0,1]$. By Brower's fixed point theorem f has fixed points. We will show that the behavior of the contagion depends crucially on the *fixed points* of this function f. Let's first give the formal definition of fixed points of f.

Definition 3.1. Given a function $f : [0,1] \to [0,1]$, c is a fixed point of $f(x)$ if and only if $f(c) = c$. Let Q_f be the set of fixed points $\{x : f(x) = x\}$.

— A fixed point c is a *stable point* if and only if there exists $\delta > 0$ such that $f(x) < x$ if $x \in (c, c+\delta]$ and $f(x) > x$ if $x \in [c-\delta, c)$. Let S_f be the set of all stable points.
— A fixed point c is a *unstable point* if and only if there exists $\delta > 0$ such that $f(x) > x$ if $x \in (c, c+\delta]$ and $f(x) < x$ if $x \in [c-\delta, c)$. U_f is defined as the set of all unstable points.
— A fixed point c of is a *touch point* if and only if $\exists d > 0, \forall x \in [0,1] : 0 < |x-c| < d, f(x) > x$ or $\forall x \in [0,1] : 0 < |x-c| < d, f(x) < x$. Let T_f be the set of touch points.

In the following we first report the detailed analysis for the case of preferential attachment graphs. In the last subsection we show how to generalize it to the case of uniform random attachment.

3.1. Main Results For Preferential Attachment

Now we are ready to state the main theorem that characterizes the behavior of general threshold contagion on *preferential attachment* graphs $PA_m(n)$.

THEOREM 3.2. *Let $\mathcal{M}_{\mathcal{G}}$ be the stochastic Markov process defined on a directed $GTC(\mathrm{PA_m(n)}, \mathrm{D}, \mathrm{I})$ contagion. The behavior of $\mathcal{M}_{\mathcal{G}}$ depends on the values of the stable fixed points of function $f(x)$ defined in Equation 2 as follows:*

(*1*) *If $f(x)$ has a unique fixed point y^* which is stable, Y_n converges to y^*. In the following three results, each subsequent result is stronger but only applicable under more restrictive settings.*

(*a*) $\forall \delta > 0$ *and* $\xi > 0$,

$$\text{Prob}[|Y_n - y^*| < \delta] = 1 - O(\frac{1}{n^{\xi}})$$

(*b*) *If* $f'(y^*) < 1$, *then* $\forall \gamma, (1 - f'(y^*))/2 > \gamma > 0$, *and* $\xi > 0$, *we have*

$$\text{Prob}[|Y_n - y^*| < O(n^{-\gamma})] = 1 - O(\frac{1}{n^{\xi}})$$

(*2*) *If* $f(x)$ *has a finite number of fixed points, then*
(*a*) $\lim_{n\to\infty} Y_n$ *exists almost surely, and* $\text{Prob}[\lim_{n\to\infty} Y_n \in Q_f] = 1$.
(*b*) $\forall s \in U_f$, $\text{Prob}[\lim_{n\to\infty} Y_n = s] = 0$.
(*c*) $\forall s \in S_f \cup T_f$, $\text{Prob}[\lim_{n\to\infty} Y_n = s] > 0$.
(*3*) *If* $f(x)$ *has an infinite number of fixed point, the process* $\{Y_i\}$ *is a martingale process and converges almost surely to some random variable* Y.

3.2. Proof of Theorem 3.2

Let's first understand the fixed points of the function f. In particular, we would like to understand the recursive structure for Y_i, i.e., the probability for a specific edge from the ith arriving node landing in an infected node. This depends on the edge attachment rule and thus needs to be done case by case. In the following we present the analysis when the selection rule is preferential.

Assume that i nodes have arrived and picked their edges. We have a total of mi edges which contribute to a total of $2mi$ degrees (including both outgoing and incoming degrees). Let I_i be the number of infected degrees (shooting from or landing on an infected node) and U_i be the number of non-infected degrees (shooting from or landing on a non-infected node). $I_i + UI_i = 2mi$ and $I_i = 2miY_i$. Given information $\mathcal{F}_i$ at time i which consists of the subgraph $\mathrm{SA_m(i)}$ and all the threshold of nodes with index smaller than i, we want to compute the value of Y_{i+1} when the $i+1$th node u_{i+1} is added, given values of Y_1 up to Y_i. For this there are three components that contribute to Y_{i+1}:

- First from previous steps we have $I_i = 2miY_i$ infected degrees.
- If the new added node u_{i+1} is infected, then the m degrees of the edges that u_{i+1} issue are infected. Thus, u_{i+1} will contribute $f(Y_i)m$ infected degree in expectation, where $f(Y_i)$ is the probability of u_{i+1} being infected.
- When u_{i+1} is added, it issues m edges to previous i nodes. Some of these neighbors are already infected, so the new edges will contribute mY_i degrees in expectation.

Define $\text{Bin}(n, p)$ as the random variable following binomial distribution, i.e., the total number of successful events out of a total of n events when each event succeeds with probability p independent of the others. Hence we get the following recurrence:

$$(2m(i+1))Y_{i+1}|\mathcal{F}_i = 2miY_i + \text{Bin}(m, Y_i) + m \cdot \text{Bin}(1, f(Y_i)),\ Y_i \in [0,1], \forall 1 \le i \le n.$$

It can be decomposed as predictable part g and noise part U

$$Y_{i+1} - Y_i|\mathcal{F}_i = \frac{1}{i+1}(g(Y_i) + U_{i+1}) \text{ for } i \ge m \tag{3}$$

$$\text{where } g(Y_i) = \frac{1}{2}(f(Y_i) - Y_i), \tag{4}$$

$$\text{and } U_{i+1} = \frac{1}{2}\left(\text{Bin}(m, Y_i)/m + \text{Bin}(1, f(Y_i)) - Y_i - f(Y_i)\right) \tag{5}$$

Define $W_i = \sum_{k=m+1}^{i} U_k/k$. Because $\mathbb{E}[U_{i+1}|\mathcal{F}_i] = 0$ and $|U_{i+1}| \leq 1$, $\{W_i : m < i \leq n\}$ is a martingale and we can rewrite the process as

$$Y_t = Y_m + \sum_{k=m+1}^{t} \frac{1}{k} g(Y_{k-1}) + W_t, \text{ for } i \geq m \tag{6}$$

3.2.1. Proof of Theorem 3.2 1a. We first analyze the case when f has a unique stable fixed point (Theorem 3.2 1a). The results of this section will be used in the multiple fixed point settings. Given an interval I of length 2δ centered at the fixed point y^* for the function f, we will show that the process will stay in the interval with probability $1 - O(1/n^{\xi})$.

Our proof has two parts. First, Lemma 3.3 shows that the noise part, W_i in Equation 6, is Cauchy-like. This says that after a sufficiently large time τ_0 the distance of two noise terms, $|W_s - W_t|$, for $s > t > \tau_0$, would be small. Second, in Lemma 3.4, given an interval I, if at certain time τ_0 the noise part is smaller than the width of the interval, then the process after $O(\tau_0)$ time will stay within I forever. The proofs of the two Lemmas are put in the full version.

LEMMA 3.3. *Given $\delta_1, \epsilon_1 > 0$, let $\tau_0 = 2\ln(1/(2\epsilon_1))/\delta_1^2$ and s, t such that $\forall s > t > \tau_0$, then* $\text{Prob}[|W_s - W_t| < \delta_1] < \epsilon_1$.

LEMMA 3.4. *Given $d > 0$, let $I = (y^* - d, y^* + d) \subset [0, 1]$ be an open interval containing the fix point y^*, and suppose there exists τ_0 such that for all s, t where $s > t > \tau_0$, then $|W_s - W_t| < d/4$. There exists $\tau_1 = O(\tau_0)$ such that $\forall k > \tau_1, Y_k \in I$.*

Combining Lemma 3.3 and 3.4 we finish the proof that the ratio will converge to any neighborhood I of the stable point with negligible probability, and that finished our proof of Theorem 3.2 1a.

3.2.2. Proof of Theorem 3.2 1b. Theorem 3.2 1b is a stronger result than Theorem 3.2 1a. In both cases the fixed point y^* is at the interior of interval $[0, 1]$. It says that when $f'(y^*) < 1$ the process $\{Y_t\}$ will converge to the fixed point y^* and also bounds the convergence rate. We decompose the process into two phases: In the first phase with high probability the process would enter and stay in the good interval I which will be defined later; in the second phase the process would approach y^* fast.

Now we define the good interval I. Recall that $f'(y^*) < 1$ and $g(y) = (f(y) - y)/2$. Define $\frac{1-f'(y^*)}{2} > \gamma > 0$. Take γ_1, γ_2 such that $\gamma < \gamma_2 < \gamma_1 < |g'(c)| \leq 1/2$. Take $d > 0$ such that

$$x \in (y^*, y^* + d], g(x) < -\gamma_1(x - y^*), \text{ and } x \in [y^* - d, y^*), g(x) > \gamma_1(y^* - x). \tag{7}$$

Now we define the good open interval I as $(y^* - d, y^* + d)$. Let

$$d_t = |y^* - Y_t|, \ \delta_t = Ad_{t_1}/t^{\gamma}$$

where $t \geq t_1$ and the value of A and t_1 will be specified later. We will prove by induction that with high probability $\forall t > t_1, d_t < C\delta_t$ for some constant $C > 0$.

Definition 3.5. We call σ is a *bad transition point* if $d_{\sigma-1} < \delta_{\sigma-1}$ and $d_\sigma \geq \delta_\sigma$, and τ is a *good transition point* if $d_{\tau-1} \geq \delta_{\tau-1}$ and $d_\tau < \delta_\tau$.

Note that by taking large enough A we can have $d_{t_1} < \delta_{t_1}$. So $d_t < \delta_t$ where $t \geq t_1$ before the first bad transition point. The following lemma shows that after certain time t_1 whenever there is a bad transition point $\sigma > t_1$ there exists a good transition point $\tau > \sigma$; moreover for all s between σ and τ the distance $d_s < C\delta_s$.

LEMMA 3.6. *Assume that* $y^* \in (0,1)$. *If* $\exists t_1$ *such that* $\forall k \geq t_1, Y_k \in I$. *Given* $\xi > 0$ *there exists* $C > 0$ *such that if there exists a bad transition point* σ *where* $\sigma > t_1$ *and* $\sigma = \Omega(\log(n))$, *there exists a good transition point* $\tau > \sigma$. *Moreover* $\forall k, d_k < C\delta_k$ *where* $\sigma < k \leq \tau$ *with probability* $1 - O(1/n^{\xi})$.

PROOF. The intuition of this proof is as follow: we run the process from time σ to $(1+\rho)\sigma$ where ρ is a small constant independent of the process, and we prove stronger claims as follows

(1) There exists t where $\sigma < t \leq (1+\rho)\sigma$ such that $d_t < \delta_{(1+\rho)\sigma} < \delta_t$.
(2) $d_k < C\delta_k$ for all k where $\sigma < k \leq \tau$ and τ is the minimum t in the first claim.

Due to space constraints we put the proofs of the two claims in the full version. With these two lemmas, we can prove the Theorem 3.2 1b: for any n, $\text{Prob}[|Y_n - y^*| < O(n^{-\gamma})] = 1 - O(1/n^{\xi})$.

For $n = t_1$ the statement of Theorem 3.2 1b is trivially true by taking proper constant A. Now, if there exists T where $t_1 < T \leq n$ such that $d_T > C\delta_T$ then there exists the first bad transition point σ_1. By Lemma 3.6 there exists the minimum good transition point $\tau_1 > \sigma_1$ and T is not between σ_1 and τ_1. Therefore we must have $t_1 < \sigma_1 < \tau_1 < T \leq n$. Inductively we find the next bad transition point σ_{i+1} and again by the Lemma 3.6 there exists a good transition point τ_{i+1} such that

$$\sigma_i < \sigma_{i+1} \text{ and } \sigma_i < T \leq n \text{ for all } i > 0.$$

Because $\{\sigma_i\}_{i>0}$ is an increasing series, T does not exists and this leads to a contradiction.

Apply union bound, all the conditions would fail with probability $O(1/n^{\xi}) + \sum_{i=t_1}^{n} O(1/n^{\xi}) \leq O(1/n^{\xi-1})$. □

3.2.3. Proof of Theorem 3.2 2 . In this case f has multiple stable fixed points. We use the framework of stochastic approximation algorithm.

Definition 3.7. A *stochastic approximation algorithm* X_n is a stochastic process taking values in $[0,1]$, adapted to the filtration $\mathcal{F}_n$, that satisfies

$$X_{i+1} - X_i | \mathcal{F}_i = \gamma_{i+1}[g(X_i) + U_{i+1}]$$

where $\gamma_n, U_n \in \mathcal{F}_n$, $g : [0,1] \to \mathbb{R}$ and the following conditions hold almost surely, (1) $c_l/n \leq \gamma_n \leq c_u/n$, (2) $|U_n| \leq K_u$, (3) $|g(X_n)| \leq K_g$, (4) $|E[\gamma_{n+1}U_{n+1}|\mathcal{F}_n]| \leq K_e\gamma_n^2$, where the constants c_l, c_u, K_u, K_g, K_e are positive real numbers.

In our problem, with filtration $\mathcal{F}_i = (Y_1, ..., Y_i)$ it satisfies

(1) $\gamma_{i+1} = 1/(i+1)$,
(2) $U_{i+1} = \frac{1}{2}(\text{Bin}(m, Y_i)/m - Y_i + \text{Bin}(1, f(Y_i)) - f(Y_i))$ is a martingale with $K_u = 4m$ and $E[\gamma_{i+1}U_{i+1}|\mathcal{F}_i] = 0$,
(3) $g(Y_i) = E[(\text{Bin}(m, Y_i) + m \cdot \text{Bin}(1, f(Y_i)) - 2mY_i)|Y_i]/2m = (f(Y_i) - Y_i)/2$ is bounded by $K_g = 1$,
(4) $|E[\gamma_{n+1}U_{n+1}|\mathcal{F}_n]| = |E[\frac{U_{i+1}}{i+1}]| = 0 \leq K_e\gamma_n^2$, where $K_e = 1$.

To prove this convergence property, we can apply the theorem in [Pemantle et al. 2007] stated as follows,

THEOREM 3.8. *If a stochastic approximation algorithm* Y_n *with continuous feedback function* g

(1) *[Corollary 2.7 in [Pemantle et al. 2007]]* $\lim_{n\to\infty} Y_n$ *exists almost surely and is in* $Q_g = \{x : g(x) = 0\}$

(2) *[Theorem 2.9 in [Pemantle et al. 2007]] Suppose there is an unstable fixed point* p *and an* $d > 0$ *such that* $\forall x : 0 < |x - p| < d$ *and* $K_l \leq E[U_{n+1}^2 | \mathcal{F}_n] \leq K_g$ *holds for some* $K_l, K_g > 0$*, whenever* $0 < |Y_n - p| < d$*.Then* $P[Y_n \to p] = 0$.

(3) *[Theorem 2.8 in [Pemantle et al. 2007]] Suppose* $p \in Q_g$ *is a stable fixed point then* $P[X_n \to p] > 0$

(4) *[Corollary 2 in [Pemantle 1991]] If* $p \in T_g$ *and* f *is differentiable,* $P[X_n \to p] > 0$

Now we can prove the convergence property,

PROOF. The first statement is a result of Theorem 3.8 (1) because g is a polynomial. The second is a result of Theorem 3.8 (1) and 3.8 (2). However to apply Theorem 3.8 (2) we have to prove $E[U_{i+1}^2]$ is bounded below by constant K_L which in our case is sufficient to prove the variance of $\mathrm{Bin}(m, Y_i) - mY_i + m\,\mathrm{Bin}(1, g(Y_i) - mg(Y_i)$ is nonzero when $0 < |Y_i - p| < d$. Formally,

$$\begin{aligned}
&\mathrm{Var}(\mathrm{Bin}(m, Y_i) - mY_i + m\,\mathrm{Bin}(1, g(Y_i) - mg(Y_i)) \\
=&\mathrm{Var}(\mathrm{Bin}(m, Y_i)) + \mathrm{Var}(m\,\mathrm{Bin}(1, g(Y_i)) + 2\,\mathrm{Cov}(\mathrm{Bin}(m, Y_i), m\,\mathrm{Bin}(1, g(Y_i)) \\
\geq& mY_i(1 - Y_i) + m^2 g(Y_i)(1 - g(Y_i)) > 0
\end{aligned}$$

The last inequality comes from $\mathrm{Cov}(\mathrm{Bin}(m, Y_i), m\,\mathrm{Bin}(1, g(Y_i)) \geq 0$ by FKG inequality. Finally, Theorem 3.8 (3) and 3.8 (4) show that Y_i will converge to arbitrary stable or touch point with positive probability. □

3.2.4. Proof of Theorem 3.2 3. In the spectial case when f has an infinite number of fixed point, because f is a polynomial with degree at most m, we have $f(x) = x$ by Fundamental Theory of Algebra. As a result, the predictable part $g(x) = 0$ in (4) and U_i is a martingale difference such that $\mathbb{E}[U_{i+1} | \mathcal{F}_i] = 0$ and $|U_{i+1} | \mathcal{F}_i| \leq 1$. Therefore our random process $\{Y_i\}$ is the martingale $Y_i = \sum_{\ell=m+1}^{i} \frac{1}{\ell}\{U_\ell\}$. To prove the convergence of martingale $\{Y_i\}$ we can use standard martingale convergence theorem (c.f. Theorem 1 in chapter 7.8 in [Grimmett and Stirzaker 2001]) to prove convergence. Because $\mathbb{E}[Y_i^2] = \sum_{\ell=m+1}^{i} \frac{1}{\ell^2} |U_\ell|^2 \leq \sum_{\ell=m+1}^{i} \frac{1}{\ell^2} < \infty$ for all i, there exists a random variable Y such that Y_i converges to Y almost surely.

3.3. General threshold cascade on stochastic-attachment graph

The analysis we did before is for the case of preferential attachment graph. Here we give the analysis for the more general case, when 1) the number of edges of the newcomer to previous nodes is sampled from a bounded distribution M with range between 1 and c_u and $E[M] = \mu_M$; 2) when the attachment rule can be either preferential or uniformly at random.

Similar to analysis in Section 3.2, we first look at the case of preferential attachment when each newcomer may choose different number of edges. Now we consider $\{Y_i : i = c_u + 1, ..., n\}$, which is a Markov process and c_u is the maximum number of edges in the distribution M. Similarly, we can compute the probability of i-th node being infected when the threshold is $R_i = k$ and $m_i = m$ edges go to previous nodes V_{i-1},

$$\mathrm{Prob}[\text{Infection of } i\text{-th node} | R_i = k, m_i = m] = \sum_{\ell=k}^{m} \binom{m}{\ell} Y_i^\ell (1 - Y_i)^{(m-\ell)}$$

Now, the probability of infection of i-th node given the forward degree $m_i = m$ is described by a function f_0^m:

$$f_0^M(Y_i) = \mathrm{Prob}[\text{Infection of } i\text{-th node}] = \sum_k \mathrm{Prob}[R_i = k] \sum_{\ell=k}^{m} \binom{m}{\ell} Y_i^\ell (1 - Y_i)^{(m-\ell)}$$

Let d_i be the total number of endpoints at time i, then $d_i + 2m_{i+1}$ would be the total number of end point at time $i+1$ if the forward degree of $(i+1)$-th node is m_{i+1}, and the recurrence relation of Y_i can be written as follows,

$$Y_{c_u+1} \in [0,1], \text{ initial condition} \tag{8}$$

$$d_{i+1}Y_{i+1}|\mathcal{F}_i = d_iY_i + \text{Bin}(m_{i+1}, Y_i) + m_{i+1}\,\text{Bin}(1, f_0^{m_{i+1}}(Y_i)) \text{ for } i > c_u + 1 \tag{9}$$

For the case of uniform random attachment when each newcomer may choose different number of edges, we define $\{Z_i : i = c_u + 1, ..., n\}$ as a Markov process, where Z_i is the ratio of infected nodes. If we define $f_1^m(Z_i)$ be the probability of i-th node being infected with m edges go to previous nodes which is

$$f_0^M(Y_i) = \sum_k \text{Prob}[R_i = k] \sum_{\ell=k}^{m} \binom{m}{\ell} Z_i^{\ell}(1-Z_i)^{(m-\ell)}$$

We can get the recursive relation as follow:

$$Z_{c_u+1} \in [0,1], \text{ initial condition} \tag{10}$$

$$(i+1)Z_{i+1}|\mathcal{F}_i = iZ_i + \text{Bin}(1, f_1^{m_i}(Z_i)) \text{ for } i > c_u + 1 \tag{11}$$

Now we are ready to state the theorem.

THEOREM 3.9. *Let $\{Y_t\}$ and $\{Z_t\}$ be the stochastic process definition above and a directed $GTC(G_{\alpha,m}(n), D, I)$ contagion.*

(*1*) *For preferential attachment, the stochastic process Y_t would converge almost surely to the stable fix point of $\frac{\mathbb{E}[Mf_0^M(y)]}{\mathbb{E}[M]}$.*

(*2*) *For uniform random attachment, the stochastic process Z_t would converge almost surely to the stable fix point of $\mathbb{E}[f_1^M(y)]$.*

Again the study of the behavior of Y_i and Z_i uses the framework of stochastic approximation algorithm 3.7.

4. UNDIRECTED PREFERENTIAL ATTACHMENT GRAPHS

In this section, we analyze the Markov process $\mathcal{M}_{\mathcal{G}}$ when the underlying network is an undirected preferential attachment graph. Here we categorize the behavior of $\mathcal{M}_{\mathcal{G}}$ based on the values of the stable fixed points of the *directed* version of $\mathcal{M}_{\mathcal{G}}$.

THEOREM 4.1. *Let $\mathcal{M}_{\mathcal{G}}$ be a stochastic Markov process defined on an undirected preferential attachment graph $\mathrm{PA_m(n)}$ with infected ratio $\tilde{Y}_n$. Suppose $f(y)$ is the function defined on a directed $\mathrm{PA_m(n)}$ in Equation 2. We have:*

(*1*) *If 1 is a fixed point of f, 0 is not a fixed point of f and the initial infected nodes $I \neq \emptyset$, then the whole network will be infected, i.e.,* $\text{Prob}[\tilde{Y}_n = 1] = 1 - o(1)$.

(*2*) *If none of 0 or 1 is a fixed point of f, then $\mathcal{M}_{\mathcal{G}}$ process will converge to a value greater than the highest stable fixed point $c = \max(S_f)$ of $\mathcal{M}_{\mathcal{G}}$ with high probability. That is,* $\text{Prob}[\tilde{Y}_n > c + m\Delta] = 1 - o(1)$ *such that $\forall\Delta$ where $0 < \Delta < (1-c)r/4$ for some r such that $(1/r)^{1/r} \geq 4\left(7mK/p^*\right)^m$, where K is the highest possible threshold of D and $p^* < \frac{1}{m(1-c)}\sum_{s=0}^{m-1}\text{Prob}[R_t = s]\sum_{\ell=s}^{m-1}\binom{m}{\ell}c^{\ell}(1-c)^{(m-\ell)}$ and c is the highest fixed point.*

In our proof, we will restrict how the contagion can proceed, which will serve to establish a lower bound for the infection ratio of the undirected contagion. First we only consider the contagion passing from low indexed nodes to high indexed nodes – just as in the directed case; then we only consider the contagion passing from high indexed nodes to low indexed nodes. We call these two processes the *forward* and *backward* processes respectively. We will repeat these processes twice. Each time, we only need to reveal the edges that can help spread an infection (i.e., the edges that point to an infected node), and the remaining edges are revealed later. We will use this to carefully manage (in)dependence so that we may employ concentration bounds.

The intuition in the analysis is the following. The first forward process is essentially the same as the directed contagion case. If there are non-zero stable fixed points then the contagion will infect a constant fraction of nodes. In fact, since the stochastic process in the directed case converges fast, among the nodes of high indices there is a good fraction of infected nodes and these infected nodes are roughly uniformly distributed. Therefore in the first backward process, these nodes will infect the nodes with small indices almost surely, which will continue to boost the propagation in the next forward round. The following analysis will make this rigorous.

Let $n_1 = C\log(n)$ and $n_2 = \mu n$ where constants C and μ will be specified later. In our first forward/backward process, we will actually only process nodes from 1 to n_2 and back to 1, but in the second round, we will process all nodes with index from 1 to n and back down to 1.

For the sake of the proof, we divide these processing steps up into three phases and presented it in the following subsections. The goal is to show that some specific property happens at the end of each phase:

(1) *First forward and backward contagion:* Run the infection in the forward direction from node 1 to the node with index n_2. Denote by I_t^F the indicator variable on whether node t is infected in the first forward process and P_t^F the probability of node t being infected, i..e, $P_t^F = \text{Prob}[I_t^F = 1]$, $t \le n_2$. Then we run the backward contagion from n_2 back to 1. Define I_t^{FB} and P_t^{FB} accordingly. We denote by Y_t^{FB} the fraction of infected nodes after the first forward and backward process for all nodes with index between 1 and t. Lemma 4.2 and 4.3 show that all the first n_1 node will be infected with high probability, i.e., $\text{Prob}[Y_{n_1}^{FB} = 1]$ is high.

(2) *Converge to highest fixed point:* Conditioned on $Y_{n_1}^{FB} = 1$, run the second forward infection to node n_2 again. We show that the infection ratio after the first n_2 nodes, denoted by $Y_{n_2}^{FBF}$, is around the highest stable fixed point c (Lemma 4.4).

(3) *Constant separation:* Conditioned on the infection ratio $Y_{n_2}^{FBF}$ being around the highest stable fixed point, c, run the infection in the forward direction from n_2 to n and backward from n to 1 to show that the infection fraction, Y_n^{FBFB} , is incremented by a constant in the second backward round (Lemma 4.5 and 4.6).

Let $M_\rho^F(s,t)$ be the event that all nodes with index within $(s,t]$ are infected with probability greater than ρ after the first forward process. Similarly define $M_\rho^{FB}(s,t)$ and $M_\rho^{FBF}(s,t)$ accordingly after the first backward process and the next forward process respectively.

4.1. First forward and backward contagion

After the first forward phase, the number of infected nodes is a constant, around one of the (non-zero) stable fixed point. The crucial part is to examine what happens in the first backward phase. We use two facts: fixing a node k, $k \le n_1$, all neighbors of node k will have a large probability to get infected, in the first forward phase (proven in Lemma 4.2); furthermore, with a fact that with high probability, early nodes have large degree (proven in the full version). Finally, we use a union bound to prove that all nodes k where $k \le n_1$ will get infected with high probability. Some of the technical proofs are put in the full version.

LEMMA 4.2 (UNIFORMITY INFECTION). *Fix a node k where $k \leq n_1$, for all $n_1 < t \leq n_2$. Let N_k^t be the event that node k is the neighbor of node t. Then there exists a constant $p^* > 0$ such that for all $\mathcal{F}_{t-1}$ we have*

$$\mathrm{Prob}[I_t^F = 1 | \mathcal{F}_{t-1}, M_\rho^F(n_1, n_2), N_k^t] \geq p^*.$$

Here recall that $\mathcal{F}_{t-1}$ is all the information at time $t-1$, i.e., the preferential attachment graph and threshold values for all nodes with index less than $t-1$, and $M_\rho^F(n_1, n_2)$ is the event that all nodes with index between n_1 and n_2 are infected with probability at least ρ after the first forward round, .

Apply Lemma 4.2 with a fact (proven in the full version) that with high probability, early nodes have large degree, the following lemma finishes the first phase by proving that $Y_{n_1}^{FB} = 1$ with high probability. Specifically, we show that every node k, $k \leq n_1$, will be infected with high probability in the first backward infection and then taking a union bound.

LEMMA 4.3. *Consider the network generated by the preferential attachment model with only the top n_2 nodes, $\mathrm{PA_m(n_2)}$, $\mathrm{Prob}[Y_{n_1}^{FB} = 1] > 1 - O(1/\log n_2)$.*

4.2. Converge to highest fixed point

Now we enter the second phase to show that the ratio, $Y_{n_2}^{FBF}$ will be around the highest stable fixed point c. We consider the following two events:

(1) Event $Y_{n_1}^{FB} = 1$;
(2) Event that $Y_{n_2}^{FBF}$ would be around the highest stable fixed point c conditioned on $Y_{n_1}^{FB} = 1$.

The intersection of these two events is what we want to prove. Since the first event happens with probability at least $1 - \varepsilon$ which is proved in Section 4.1, we now show a bound on the second event.

LEMMA 4.4. *Conditioned on $Y_{n_1}^{FB} = 1$, $\forall \delta, \xi > 0$, and c being the highest stable fixed point,*

$$\mathrm{Prob}[|c - Y_{n_2}^{FBF}| < \delta] > 1 - O(\frac{1}{n_2^\xi}) = 1 - O(\frac{1}{n^\xi}).$$

PROOF. Similar to Lemma 3.4 we take $n_1 = C\log(n)$ where C is large enough such that at time $t \geq n_1$ the step size $|Y_{t+1} - Y_t|$ is smaller than $\delta/4$. Because 1 is not a fixed point, $f(x) < x$ where $c - \delta \leq x \leq 1$, for $n_1 < t < n_2$ the predictable part $g(Y_t)$ will push the ratio Y_t into the interval $(c - \delta, c + \delta)$ and a similar argument in Lemma 3.4 shows that Y_t will stay there with high probability $1 - O(\frac{1}{n^\xi})$. □

4.3. Constant separation phase

Finally, for third phase, we reveal the edges from node n_2 to n, and show that the infected ratio Y_n^{FBFB} after the second backward contagion will have a constant improvement, i.e., $Y_n^{FBFB} > c + m\Delta$ where $\Delta > 0$ is independent of n. Let I_t^{FBFB} denote the indicator function that node t is not infected during the second forward infection but getting infected in the second backward infection. Because each additional infection of a node, I_t^{FBFB} would contribute at least m to weighted infection ratio Y, and it's sufficient to show that

$$\mathrm{Prob}[\sum_{k=1}^{n} I_k^{FBFB} > \Delta n] > 1 - \eta. \tag{12}$$

We use second moment argument to prove it. Lemma 4.5 shows the expected increment ratio $Y_n^{FBFB} - Y_n^{FBF}$ is greater than some constant $m\Delta$, that is sufficient to have $E[\sum_{k=1}^{n} I_k^{FBFB}] > \Delta n$. And the second lemma 4.6 shows that the variance of $\sum_{k=1}^{n} I_k^{FBFB}$ is small.

Let $M^{FBF}_{(c-\delta,c+\delta)}(n_2, n)$ be the event that $\forall t$ where $n_2 < t \leq n$, $Y_t^{FBF} \in (c - \delta, c + \delta]$.

LEMMA 4.5. *If* $n_2 = rn$, *then* $\exists \Delta > 0$ *s.t.* $E[\sum_{k=1}^{n_2} I_k^{FBFB} | M^{FBF}_{(c-\delta,c+\delta)}(n_2, n)] \geq \Delta n$.

The following lemma show that the variance of $\sum_{k=1}^{n_2} I_k^{FBFB}$ is small.

LEMMA 4.6. $\mathrm{Var}[\sum_{k=1}^{n_2} I_k^{FBFB} | M^{FBF}_{(c-\delta,c+\delta)}(n_2, n)] = O(n)$.

Apply these two lemmas we have $\mathrm{Prob}[\sum_{k=1}^{n_2} I_k^{FBFB} | M^{FBF}_{(c-\delta,c+\delta)}(n_2, n)]] > 1 - o(1)$, and $\mathrm{Prob}[\sum_{k=1}^{n_2} I_k^{FBFB}] > 1 - o(1)$, since $\mathrm{Prob}[M^{FBF}_{(c-\delta,c+\delta)}(n_2, n)]] > 1 - o(1)$. Combine these lemmas we can prove Theorem 4.1:

PROOF. The first part of the proof is derived from Lemma 4.3 since $Y_{n_1} = 1$ with high probability and $f(1) = 1$ is a fixed point, then all the nodes after n_1 will get infected and $\mathrm{Prob}[Y_n = 1] = \mathrm{Prob}[Y_{n_1} = 1] = 1 - o(1)$. In second part, the event that $\tilde{Y_n} > c + m\Delta$ holds if the following is true.

(1) The infected ration $Y_{n_2}^{FBFB}$ is around highest stable point c c.f. Lemma 4.4;
(2) The increment is greater than constant c.f. (12).

By union bound, the event fail with probability less than $1 - o(1)$. □

5. SIMULATIONS

We ran simulations on model networks and real world data sets to understand the behavior of a general threshold contagion and its dependency on threshold distribution D, the network structure, and the selection of initial seeds.

Model networks We generate graphs using the stochastic attachment model and run a contagion in both the directed and undirected version. We use two threshold distributions D_1 and D_2. In D_1, the probability of taking a threshold of $1, 2, 7$ is $0.22, 0.39, 0.49$ respectively; in D_2, the probability of taking a threshold of $1, 2, 5, 7$ is $0.1, 0.4, 0.45, 0.05$ respectively. Using definition of function f in Equation 2, with $m = 5$ and D_1, f has one fixed point equal to 0.558.With $m = 6$ and D_2, f has two fixed points 0.875 and 0.521. In each run of the simulation, we vary I to be a fraction $\beta \in [0, 1]$ of a constant number of the first 6 nodes for D_1 and first 7 nodes for D_2 of the network.

Directed network. We create a network G_1 based on Definition 2.2, in which each newcomer choose m edges that are preferentially attached to earlier nodes. G_1 is directed, each edge pointing from a high indexed node to a low indexed node. Figure 1 show the results of running a contagion over G_1 using D_1 and D_2 with different sets of seeds. For different runs, the ratio of infected nodes converges to one of the stable fixed points. When f has multiple fixed points (as in the case of D_2), the way that the first few nodes are infected typically determine the infection rate of the entire network.

Undirected network. We take G_1, make all edges undirected, call it G_2. Then we run contagion in alternating forward and backward steps. See Figure 2. The first forward step behaves the same way as contagion on the directed network. The first backward step uniformly infects more nodes everywhere. In the case of D_2, the next foward phase infects a large number of nodes. Additional steps do not change the infection state much.

DBLP and Web graphs We use two real world networks: the Stanford web graph (a directed network) and the DBLP co-authorship network (an undirected network).

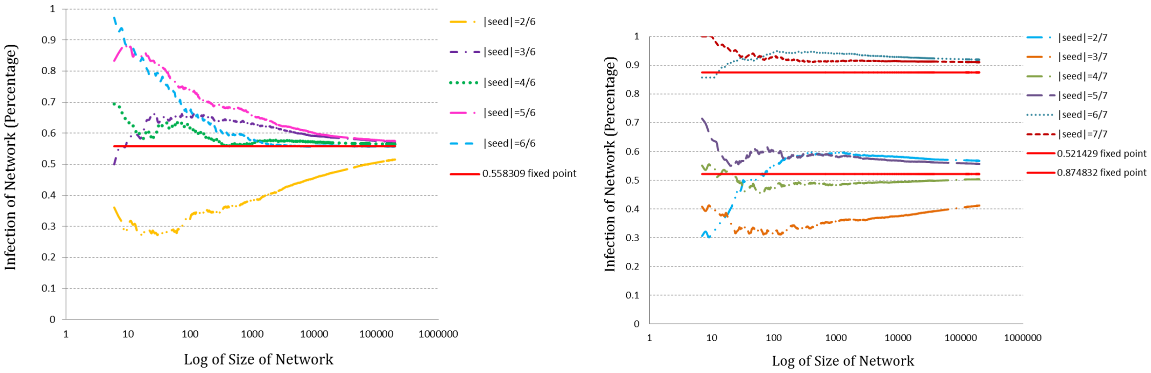

Fig. 1. Contagion using threshold distribution D_1 in (Left) and D_2 in (Right) with different initial seeds on the directed preferential attachment graph.

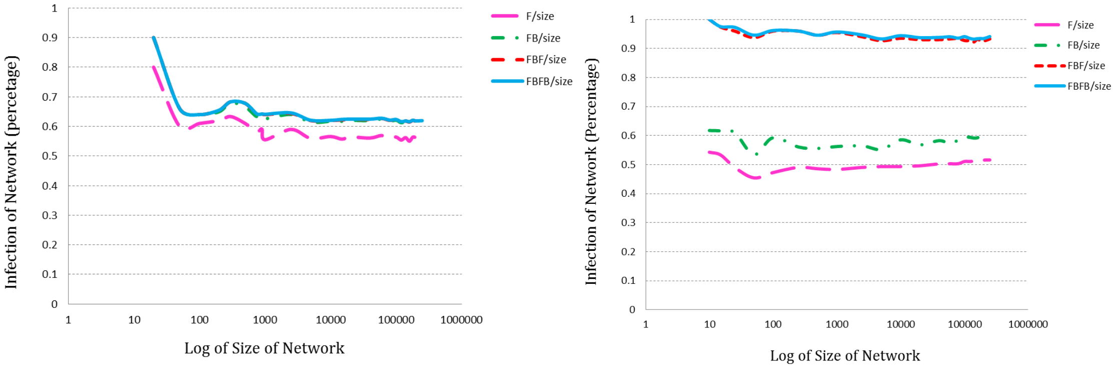

Fig. 2. Contagion using threshold distribution D_1 in (Left) and D_2 in (Right) with different initial seeds on undirected preferential attachment graph.

(1) Stanford web graph: Each node represents a page from Stanford University (stanford.edu) and there is a directed edge from u to v if u has a hyperlink to v. The network contains $281,903$ nodes and $2,312,497$ edges.

(2) DBLP co-authorship network: The nodes are authors and there is an undirected between two nodes if they have published at least one paper together. This data set has $317,080$ nodes and $1,049,866$ edges.

To understand contagion on real networks, we first try to fit our stochastic attachment graph model. For that, we generate an arriving order from the real world graphs. There can be multiple ways to do so. Here we iteratively remove the lowest degree node, with ties broken arbitrarily. Then we take the reversed order and use it as the arriving order of the nodes. If the network is directed, we iteratively remove the node of lowest in-degree. Next, each node v has a degree d_v referring to the number of edges to the lower indexed nodes. We collect all such degrees d_v, $\forall v$, and use it for the outgoing degree distribution M. Then we generate a network G' using the stochastic attachment model with outgoing degree distribution M. Here, we set the number of nodes of the network to be $300,000$, which is almost the same as the number of nodes in both Stanford and DBLP data sets. We create a complete graph of m nodes, where m is the expectation of the outgoing degree distribution M, which is 6 for the Stanford data set and 3 for the DBLP data set. For the attachment rule we introduce a parameter $\alpha \in [0,1]$ as the probability that an edge is attached using the preferential rule. If $\alpha = 0$ all edges are attached uniformly at random; if $\alpha = 1$ all edges are attached preferentially. In experiments, we use $\alpha = 0, 0.25, 0.5, 0.75, 1$.

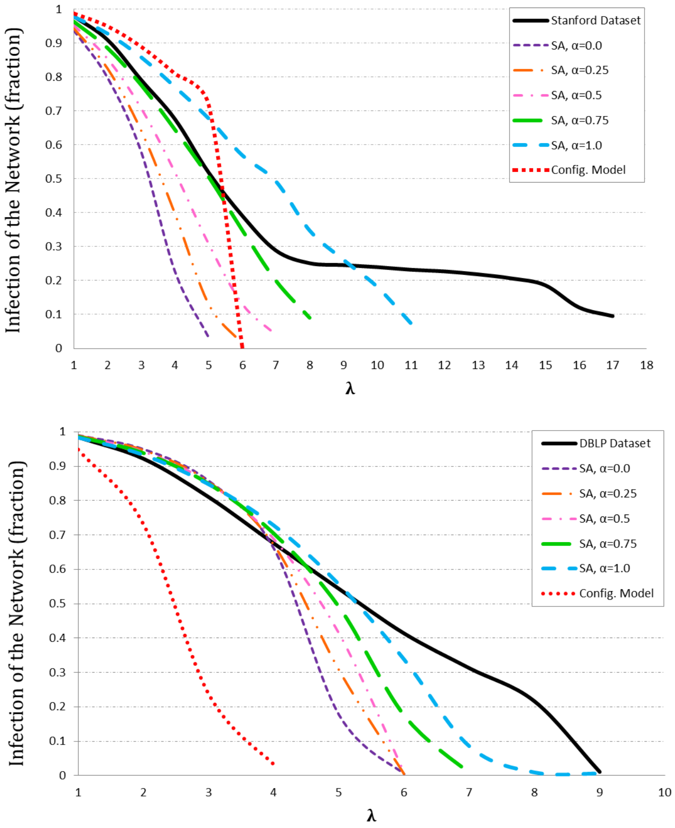

Fig. 3. Contagion on (Top) Stanford web graph and (Bottom) the DBLP coauthorship graph, stochastic attachment models and configuration models.

For contagion model, we take two approaches. First, we take D, the threshold distribution to be the Poisson distribution with parameter λ. We start each of the experiments from $\lambda = 1$ and increase its value until the total infection rate of the network drops below 1%. Second, we run a k-complex contagion model, in which all nodes have threshold k. We take seeds as the 25 lowest indexed nodes.

We run these two contagions over both real networks and their corresponding generated model networks. For comparison, we also generate a network using the configuration model following the same degree distribution of the real world network.

Figure 3 shows the results where the threshold distribution is a Poisson one. It can be observed that the behavior of contagion on the generated stochastic attachment graph (especially the one with $\alpha = .75$) matches the behavior of the real world graph fairly well, while the configuration model (though having the same degree distribution) does so poorly. Tables I and II show the confidence intervals for the most similar model for DBLP dataset, which is SA with $\alpha = 1$, and for Stanford dataset, which is SA with $\alpha = 0.75$, under different values of λ.

Figure 4 shows the results for k-complex contagion. Our models, though with infection rate shifted away from the behavior of the real world graph, is still much better than the behavior of configuration model (for which the infection rate is zero for any k complex contagions, $k \geq 2$). In particular, we believe this is partly due to the lack of community structures in the configuration model.

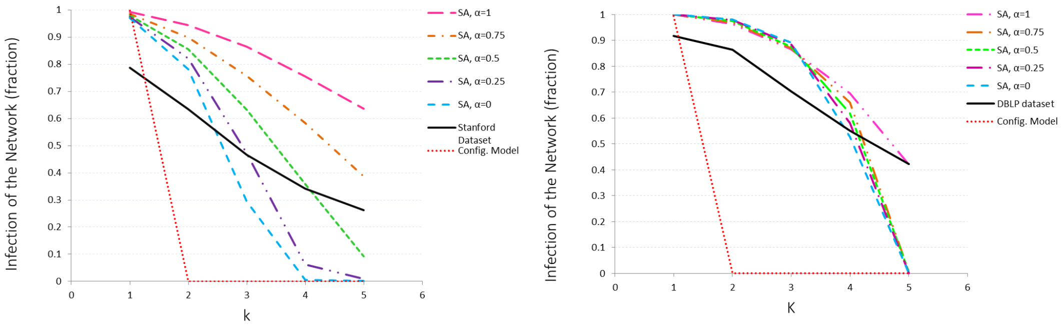

Fig. 4. Contagion on Stanford web graph (left) and the DBLP coauthorship graph (right) under stochastic attachment models and configuration models.

$\alpha = 1$	Mean	95% CI
$\lambda = 1$	.9835	[.9832,.9837]
$\lambda = 2$	.9334	[.9338,.9329]
$\lambda = 3$	.8486	[.8478,.8498]
$\lambda = 4$	.7260	[.7220,.7283]
$\lambda = 5$	.5608	[.5564,.5662]
$\lambda = 6$	.3105	[.3004,.3488]
$\lambda = 7$	.0588	[.0274,.0855]
$\lambda = 8$	.0113	[.0061,.0294]

Table I.

$\alpha = .75$	Mean	95% CI
$\lambda = 1$	.9631	[.9611,.9663]
$\lambda = 2$	.8853	[.8834,.8886]
$\lambda = 3$	.7762	[.7760,.7789]
$\lambda = 4$	.6451	[.6198,.6798]
$\lambda = 5$	.5058	[.4872,.5232]
$\lambda = 6$	.3487	[.3212,.3623]
$\lambda = 7$	.1997	[.1712,.2198]
$\lambda = 8$	.0903	[.0672,.1271]

Table II.

6. CONCLUSION

This paper initiates the study of complex contagion with general thresholds. One takeaway is that stochastic attachment graph model can be used to estimate the behavior of contagion on real data sets better than configuration models.

References

Lada A Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd International Workshop on Link discovery*. 36–43.

Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. 2006. Group Formation in Large Social Networks: Membership, Growth, and Evolution. In *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 44–54.

Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. 2013. The diffusion of microfinance. *Science* 341, 6144 (2013).

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

J. Coleman, E. Katz, and H. Menzel. 1957. The diffusion of an innovation among physicians. *Sociometry* 20 (1957), 253–270.

James S. Coleman, Elihu Katz, and Herbert Menzel. 1966. *Medical Innovation: A Diffusion Study*. Bobbs-Merrill Co.

Devdatt P Dubhashi and Alessandro Panconesi. 2009. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press.

D.Watts, P.Dodds, and M.Newman. 2002. Identity and Search in Social Networks. *Science* 296 (2002), 1302–1305.

D.Watts and S.Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393, 6684 (1998), 409–410.

Roozbeh Ebrahimi, Jie Gao, Golnaz Ghasemiesfeh, and Grant Schoenebeck. 2014. How Complex Contagions Spread Quickly in the Preferential Attachment Model and Other Time-Evolving Networks. *arXiv preprint arXiv:1404.2668* (2014).

Roozbeh Ebrahimi, Jie Gao, Golnaz Ghasemiesfeh, and Grant Schoenebeck. 2015. Complex

Contagions in Kleinberg's Small World Model. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science (ITCS)*. 63–72.

Golnaz Ghasemiesfeh, Roozbeh Ebrahimi, and Jie Gao. 2013. Complex contagion and the weakness of long ties in social networks: revisited. In *Proceedings of the fourteenth ACM conference on Electronic Commerce*. 507–524.

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

M. Granovetter. 1978. Threshold Models of Collective Behavior. *Am. Journal of Sociology* 83, 6 (1978), 1420–1443.

Geoffrey Grimmett and David Stirzaker. 2001. *Probability and random processes*. Oxford university press.

Matthew O. Jackson. 2008. *Social and Economic Networks*. Princeton University Press, Princeton, NJ, USA.

David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 137–146.

Jon Kleinberg. 2000. The small-world phenomenon: an algorithm perspective. In *Proceedings of the 32-nd annual ACM symposium on Theory of Computing*. 163–170.

Jon M. Kleinberg. 2001. Small-World Phenomena and the Dynamics of Information.. In *NIPS*. 431–438.

Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S. Tomkins. 1999. The web as a graph: measurements, models, and methods. In *Proceedings of the 5th annual international conference on Computing and combinatorics*. 1–17.

R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. 2000. Stochastic models for the Web graph. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS '00)*. 57–.

Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. 1999. Extracting Large-Scale Knowledge Bases from the Web. In *Proceedings of the 25th International Conference on Very Large Data Bases*. 639–650.

J S Macdonald and L D Macdonald. 1964. Chain Migration, Ethnic Neighborhood Formation and Social Networks. *The Milbank Memorial Fund Quarterly* 42, 1 (1964), 82–97.

Robin Mermelstein, Sheldon Cohen, Edward Lichtenstein, John S Baer, and Tom Kamarck. 1986. Social support and smoking cessation and maintenance. *Journal of consulting and clinical psychology* 54, 4 (1986), 447.

Elchanan Mossel and Sebastien Roch. 2007. On the submodularity of influence in social networks. In *Proc. of the thirty-ninth annual ACM symposium on Theory of computing*. 128–134.

Elchanan Mossel and Sébastien Roch. 2010. Submodularity of Influence in Social Networks: From Local to Global. *SIAM J. Comput.* 39, 6 (2010), 2176–2188.

M E J Newman and D J Watts. 1999. Scaling and percolation in the small-world network model. *Physical Review E* 60, 6 (1999), 7332–7342.

Robin Pemantle. 1991. When are Touchpoints Limits For Generalized Polya Urns. In *Proceedings of the American Mathematical Society*, Vol. 113. 235–243.

Robin Pemantle and others. 2007. A survey of random processes with reinforcement. *Probab. Surv* 4, 0 (2007), 1–79.

Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*. 695–704.

J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. 2012. Structural Diversity in Social Contagion. *Proc. National Academy of Sciences* 109, 16 (April 2012), 5962–5966.