INTERNATIONAL JOURNAL OF INTELLIGENT CONTROL AND SYSTEMS
VOL. 14, NO. 1, MARCH 2009, 87-96

A History Sensitive Diffusion Network: Preliminary Model and Simulation

Yu ZHANG, Jiang WU and Bin HU

Abstract- **Diffusion is a process by which information, viruses, ideas and new behavior spread over social networks. The traditional *independent cascade model* gives activated nodes a one-time chance to activate each of its neighboring nodes with some probability. This paper extends the traditional model to an iterative fashion. We propose a model called the *iterative cascade model* that allows activated nodes to receive more than a one-time chance to activate their neighbors. We formally define a deterministic model for calculating the probability of activity for any arbitrary node at any arbitrary time in the iterative cascade model. We also perform an extensive empirical study on our model under different network settings and topological structures. This empirical study covers common problems in the cascade model, such as maximization, outbreak detection, and submodularity. In addition, we evaluate our model in a real-world social network in viral advertising. These simulations have showed its power to observe and explain the emergent phenomena in the macro level when changing parameters in the micro level.**

***Index Terms*— Social Network, Information Cascade**

1. INTRODUCTION

Diffusion is a process by which information, viruses, ideas and new behavior spread over social networks [15]. The research of diffusion in social networks originated from Rogers' Innovation Decision Process Theory in 1962 [18]. He formalized that innovations would spread through society in an S curve, as the early adopters select the technology first, followed by the majority, until a technology or innovation is common. This work was extended in Granovetter's treatment of the phenomenon of collective behavior in 1978 [12]. He introduces a threshold model and uses it to examine the occurrence of riots and their perceived domino-effect growth pattern. The result was an early threshold model for collective behavior. In present day, diffusion models of social networks have been studied in a variety of fields ranging from epidemiology [16], to marketing [17], to technology transfers [4, 5], to computer virus transmission [2], and to power systems [22].

As described by Kempe et. al. [14], there are two basic diffusion models: 1) the *linear threshold model* [12], in which a node becomes active if a predetermined fraction, called a threshold, of the node's neighbors are active, and 2) the *independent cascade model* [8], whenever a node becomes active, it gets a one-time chance to activate each of its neighboring nodes with some probability. In both models, influence maximization problems are typically phrased in the following terms [9]: given some value *k* and some diffusion network with a set of nodes *N*, the goal is to select an initially active *k*-node subset from *N*, such that the number of nodes in *N* that eventually becomes active is maximized. It has been showed that many objective functions for the maximization problem is submodular [14], that is, beyond some point (it is called the tipping point [10]), adding more nodes to *k* yields less and less additional diffusion power to the network. The maximization problem is NP-Hard [14] but the answer can be approximated by heuristics like greedy [19], hill climbing [14], and simulated annealing [13], and can be bounded on the quality of the cost-effective outbreak detected in the network [15].

There is a rich literature in both diffusion models; especially the independent cascade model has gained much attention in present day. Goldenberg et al. [11] simulate Word-of-Mouth information diffusion through strong ties among members of the same network and weak ties among individuals belonging to different network. They found the influence of weak ties on the information diffusion is almost as strong as the influence of strong ties. Cowan and Jonard [7] study diffusion of knowledge in different network structures. They find that the performance of the system exhibits clear "small world" properties, in that the steady-state level of average knowledge is maximal when the structure is a small world (that is, when most connections are local, but roughly 10 percent of them are long distance). In viral marketing and advertising, Leskovec et. al [15] simulate information cascade in a real person-to-person recommendation network. They discover that the distribution of cascade sizes is approximately heavy-tailed; cascades tend to be shallow, but occasional large bursts of propagation can occur.

The model that we propose is called the *Iterative Cascade Model*. It can be understood as a modified Independent Cascade Model, but is different from the generalized framework (proposed by Kempe) designed to unify the Independent Cascade Model and the Linear Threshold Model underneath a single mathematical threshold. For one major difference, our model allows that activated nodes receive more than a one-time chance to activate their neighbors. A node can switch from being uninformed to being informed, but not the reverse. The rationale of this model is the classical threshold mechanism of collective action: a consumer does not feel

Manuscript received January 1, 2009. This work was supported in part by the U.S. National Science Foundation under Grants IIS 0755405 and CNS 0821585, and the China National Natural Science Fund No. 70671048, 70731001. This paper is extended from "A History Sensitive Cascade Model in Diffusion Networks" published at *Spring Simulation Multiconference 2009, Agent-Directed Simulation Symposium*, San Diego, CA, USA, March, 2009.

Y. Zhang is with Dept. of Computer Science, Trinity University, One Trinity Place, San Antonio, TX 78212, USA (e-mail: yzhang@cs.trinity.edu); J. Wu and B. Hu are with Huazhong University of Science and Technology, China (e-mail: jiangwu.john@gmail.com and bin_hu@mail.hust.edu.cn).

social pressure if just a few people around her behave in a particular way but once these people reach a certain number then she suddenly decide to change her mind and she behaves differently [12]. Iterative interactions between people are often observed in real world. For example, in viral marketing and advertising, a customer may not decide to buy a product at their first time receiving the recommendation or watching the advertisement, but they may gradually accept the product and decide to buy it after several rounds of such interactions.

This paper formally defines the iterative cascade model and performs an extensive empirical study to it. One can easily see that in the iterative cascade model, the probability that a node is active (an adopter of the innovation) will be reinforced in each interaction and approach 1.0 finally. But we think the common problems in the generalized framework, such as maximization, outbreak detection, and submodularity, are all worthy of study in the iterative cascade model. So we propose a deterministic model for calculating the probability of activity for any arbitrary node at any arbitrary time. We first define the probability model for a tree-structured graph. When extending this to a general graph with cycles, we propose a Markov model for calculating the probability. Next we perform an extensive empirical study on our model under different network settings and topological structures. Finally, we evaluate our model in a real-world social network in viral marketing. These formalized model and empirical study can be a solid foundation for further analyzing and explaining the emergent phenomena in the macro level when changing parameters in the micro level in diffusion social networks.

2. THE MODEL FOR A TREE-STRUCTURED GRAPH

Let $G = (V,E)$ where V is a set of vertices and E is a set of edges. Each vertex v in V corresponds to a Boolean value $A(v)$, denoting whether the vertex is "active" or not. Each edge $e_{v,u}$ (leaving v and entering u) is weighted with a value $0 < W(e_{v,u}) <= 1$, representing the probability that if v is active in time step $t = k$, then u will be active in time step $t = k+1$. $W(e_{v,u})$ is called the spreading probability. Let $targets(v)$ represent all u such that $e_{v,u}$ is in E. The activity value of vertices is updated as the Simulation Algorithm in Fig. 1.

```
For time step = 1 to k
  For each vertex v in V
    If A(v) = true
      For each vertex u in targets(v)
        threshold = a random number between 0
and 1;
        If threshold < W(e_{v,u})
          Set A(u) = true;
```

Fig. 1 Simulation Algorithm

The question we wish to answer is, given some number k and some vertex v, what is the probability that v will be active at $t = k$? We assume (for the sake of eschewing trivialities) that at time step $t = 0$, there exists at least one node a such that $A(a) = true$. Or, as we may alternatively posit, $P(a_0) = 1$, meaning simply that the probability of a being active on the 0^{th} time step is 1. If this were not the case, then there would be no activity for any nodes on any time step. The proof for this can easily be extrapolated from the above algorithm, whose only alterations to the state of the system occur within an If-statement, the condition of which is that $A(v)$ be *true*. Thus, a system starting out with $A(v)$ being *false* for all v will never change states.

We now go about examining properties of this model in order to derive a method for calculating the probability of activity for any arbitrary node at any arbitrary time. Throughout the discussion, suppose for simplicity's sake that we have only one active vertex a at time $t = 0$. Ultimately, we will drop this assumption.

Theorem 1 (Definition of Inactivity): If $P(x_k) = 0$, then $P(x_{k-i}) = 0$, for all $0 <= i <= k$.

Proof: That is equivalent to asserting that, if a vertex is not active during $t = k$, it was never active at a previous time step. The proof follows from the fact that our previous description of the model does not provide for any conditions in which an active vertex might deactivate. So an inactive vertex was never active previously. □

Calculating the probability the neighbor of an active vertex will be active in the next time step is a simple matter.

Lemma 1: Assuming that 1) x only has one edge coming to it from an active vertex (which we call a*), 2) $P(x_0) = 0$, 3) $P(a_0) = 1$, and 4) $t = 1$,*

$$P(x_1) = W(e_{a,x}). \quad (1)$$

Proof: The proof follows from the definition of the model. In time step $t = 0$, vertex a is active and, thus, will activate any inactive neighbor n with a probability equal to the weight of the edge from a to n. Since this lemma assumes that $t = 1$ and that x was not active at $t = 0$, we know that x has had no chance to become active before $t = 1$. Thus, $P(x_0) = 0$ and $P(x_1) = W(e_{a,x})$, as can be seen from the statement of the Simulation Algorithm. Matters, however, will not be so simple in the following case. □

Lemma 2: Dropping the forth assumption above, assume that 1) x has only one edge coming to it from an active vertex (which we call a), 2) $P(x_0) = 0$, 3) $P(a_0) = 1$.

$$P(x_{k+1}) = [1 - P(x_k)]W(e_{a,x}) + P(x_k). \quad (2)$$

Proof: It helps to note that, because of the first assumption above (combined with the Definitions of Activity and Inactivity), x has been the neighbor of one and only one active vertex ever since $t = 0$, i.e. $P(a_k) = 1$ for all k. This is useful because it means that we need only pay attention to the following two ways that x might become active in time step $t = k + 1$.

1) x had been activated by a already during some time step $t <= k$, or
2) x is activated by a during the step $t = k + 1$.

We can represent the first of these cases like so,

$$already_{k+1} = P(x_k). \quad (3)$$

And the second like so,

$$during_{k+1} = [1 - P(x_k)]W(e_{a,x}). \quad (4)$$

This equation makes use of the fact that $P(x_k)$ is the probability that x was active before $t = k+1$, and thus, $1 - P(x_k)$ is the probability that x was not active before $t = k + 1$. Since the probability of $A(x)$ being true already at $t = k+2$ and of becoming true during $t = k + 2$ are disjoint, the probability of x being active at $t = k + 2$ is given by a simple sum:

$$P(x_{k+1}) = during_{k+1} + already_{k+1} \quad (5)$$

which is the same as

$$P(x_{k+1}) = [1 - P(x_k)]W(e_{a,x}) + P(x_k) \quad \square$$

Lemma 3: Dropping, now, the third assumption above, we need not assume that $P(a_k) = 1$ for all k. Instead, we assume that 1) x has only one edge coming to it from any potentially active vertex a, and 2) $P(x_0) = 0$. In this case, the activity probability is:

$$P(x_{k+1}) = [P(a_k) - P(x_k)]W(e_{a,x}) + P(x_k) \quad (6)$$

Proof: The proof is almost identical to the one above except that the probability of x being activated during $t = k+1$ is this:

$$during_{k+1} = [P(a_k) - P(x_k)]W(e_{a,x}). \quad (7)$$

The justification for the term $P(a_k) - P(x_k)$ is again similar to the previous proof. It makes use of the fact that if x has ever had an active influencer, the probability of x being active is completely dependent on the probability of a being active, because a is and always has been x's only means of becoming active. Finally,

$$P(x_{k+1}) =[P(a_k) - P(x_k)]W(e_{a,x}) + P(x_k) \quad \square$$

Using the above lemmas, we can calculate the activity probability of any arbitrary vertex on any arbitrary time step, provided that the graph is a tree and that no vertex has more than one incoming edge.

Theorem 3: Let T = (V,E) be a graph without cycles, and let there be no two edges in E, $e_{w,x}$ and $e_{y,z}$, such that x = z. We will use the function influencer(u) to denote the vertex v such that targets(v) includes u.

$$P(v_{k+1}) = \begin{cases} 1, & \text{if } P(v_0) = 1 \\ [P(\inf luencer(v)_k) - P(v_k)] \times W(e_{\inf luencer(v),v}) + P(v_k), & otherwise \end{cases} \quad (8)$$

Proof*:* The first part of the formula follows from the definition of activity. The second part follows from observing that the vertices will always satisfy the assumptions of either Lemma 2 or 3:

1) Vertex v will always have only one potentially active influencer.
2) $P(v_0)$ will, in all cases be 0, or else the first part of the formula would have been applicable. $\square$

3. THE MODEL FOR A GENERAL GRAPH

Expanding our ability to perform this calculation on graphs with cycles presents several problems. The tip of the iceberg with regard to cycles can be illustrated by the situation in which $P(v_k) > 0$ and v has an edge to u, which is inactive on time step $t = k$. Also suppose that u has an edge to v and no edges to or from any other vertices. On time step $t = k+1$, Lemma 3 tells us that $P(u_{k+1})$ will equal $P(v_k) \times W(e_{v,k})$. Then, in time step $t = k+2$, the vertex v would potentially have an extra active influencer, u, but this active influencer should NOT alter $P(v_{k+2})$ because $P(u_i)$ is, for all i, dependent on $P(v_{k+2})$, a consequence of u having no influencers aside from v. This example is just a cycle of length two. Much more complicated cases arise when cycles are longer. The underlying problem can be grasped by realizing that the system we are modeling is based on information of some type being passed between vertices, and these vertices cannot rely on information they have not yet received. So cycles allow a vertex v to receive information M that it was originally responsible for relaying – a case that should not raise v's probability of having previously received M. We now present a Markov Model approach for ascertaining the probability of any vertex being active at any time step in a general graph.

First, we consider the graph to be a finite state system, where "state" is understood as some combination of activated vertices in V. The state of the graph can be represented as a Boolean-valued "activity" vector a of length n, where n is the size of V. Let each vertex be arbitrarily labeled with a unique value between 0 and n-1, and let this label value be denoted $L(v)$. A value of 1 in the index i (from the right) of the vector a indicates that the vertex v, such that $L(v) = i$, is active. More compactly, $a[L(v)] = A(v)$. Thus, the vector

$$a = [1\ 0\ 1\ 0\ 0\ 1]$$

indicates that for all v such that $L(v) = 0, 3,$ or 5, $A(v) = true$. Now, let us use another binary-valued vector called s to denote every possible binary number that the vector a can represent. We call s the "state" vector and it will be of length 2^n. s will have only a single 1-value; the rest will be 0s. The index of s' 1-value, when converted to base two will be the binary number represented by a corresponding activity vector a. Thus, the state space vectors

$s = [1\ 0]$ $s = [0\ 1]$
$s = [1\ 0\ 0\ 0]$ $s = [0\ 1\ 0\ 0]$ $s = [0\ 0\ 1\ 0]$ $s = [0\ 0\ 0\ 1]$

would map to the following activity vectors respectively:

$a = [\ 0\ ]$ $a = [\ 1\ ]$
$a = [0\ 0]$ $a = [1\ 0]$ $a = [0\ 1]$ $a = [1\ 1]$

We are introducing the *s* vector in order to serve as the initial state vector for a Markov chain that we will build from *G*, allowing us to calculate the exact probability of some vertex being active at some time step. Furthermore, each binary value represented by the list *a* can be considered an index into the matrix we will construct. For example, suppose we have the vertex set *V = {0,1,2}*, where *G* is fully connected and the weight of all edges shall be set to 0.5. Our transition matrix would look like this in Fig. 2 (with columns and rows labeled):

	[]	***0***	***1***	***0,1***	***2***	***2,0***	***2,1***	***2,1,0***
[]	*1.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*
0	*0.0*	*0.25*	*0.0*	*0.25*	*0.0*	*0.25*	*0.0*	*0.25*
1	*0.0*	*0.0*	*0.2*	*0.25*	*0.0*	*0.0*	*0.25*	*0.25*
0,1	*0.0*	*0.0*	*0.0*	*0.25*	*0.0*	*0.0*	*0.0*	*0.75*
2	*0.0*	*0.0*	*0.0*	*0.0*	*0.25*	*0.25*	*0.25*	*0.25*
2,0	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.25*	*0.0*	*0.75*
2,1	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.25*	*0.75*
2,1,0	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*1.0*

Fig. 2 The Transition Matrix

Arranging the combinations of vertices in the orders given will help us later to collect the necessary information for ascertaining the total probability of a given node being active. For the moment, we give the algorithm in Fig. 3 for constructing the above matrix for a general graph *G = (V,E)*.

Upon obtaining the stochastic state transition matrix *A*, the calculation of each possible state reachable from an initial binary state vector *s* – containing only a single 1-value, can be obtained as follows:

$$p = s \times A^k \tag{9}$$

where *k* is the time step for which you want a value. The resulting vector *p* has the property that *p[i]* equals the probability that *s[i] = 1* at time step *t = k*.

To find $P(v_k)$ for an arbitrary vertex *v*, we simply need to sum the elements of *p* whose state involves *v* being active:

$$P(v_k) = s \times A^k \times AllRelevantStates(L(v)) \tag{10}$$

where *AllRelevantStates(L(v))* merely returns a vector *x* such that *x[j]* is 1 if the binary representation of j contains a 1 in the *L(v)*'s location. Otherwise *x[j]* is a 0. For example, if $|V| = 3$, then we can use the following familiar pattern to discover all states in s that involve a given vertex v with label *L(v)*.

000 001 010 011 100 101 110 111

where the column to the far right is the vector returned by *AllRelevantStates(L(v))* for the vertex with label 0. The middle column will be returned for the vertex labeled 1. And the left hand column will be returned for the vertex labeled 2.

So in the end, $P(v_k)$ can be written:

$$P(v_k) = p \times AllRelevantStates(L(v)). \tag{11}$$

This sums up the values of *p* that refer to the probability of *G* being in a state where *v* is active. The reason we can sum these probabilities is that each state whose transition probability we calculate during the construction of *a* is disjoint, as can be seen from the portion of the algorithm labeled *Make Probabilities Disjoint*, which calculates, for every initial state *S*, the probability of activating any subset of all vertices *a* that could be activated by vertices in *S*, and (most importantly) the probability of *not* activating any other vertices in *a*. Thus, state transition probabilities do not "overlap," so to speak.

```
Function stochasticMatrix()
  For Each (S in Powerset(V))
    Let A = All vertices activatible by set S;
    Let E = edges from S to A;
    Let P be an empty hash structure from vertex keys to
      probability values;
    For Each (target in E)
      W = [weights of edges going to target];
      P += {target => combinedProbability(W)};
    //Make Probabilities Disjoint:
    For Each (S2 in Powerset(A))
      P2 = P;
      For Each (p in P2)
          If (Key(p) is in A && Key(p) is not in S2)
              Let Value(p) = 1–p;
      Matrix[indexOf(S)][indexOf(S union S2)] = Product p

Function combinedProbability(W)
  Let answer = 1
  For Each w in W
      answer = answer×(1–W);
  Return 1–answer;

Function indexOf(S)
  Let answer = a list of zeros of length max(S);
  For Each v in S
      answer[L(v)] = 1;
  Return decimal value of the binary number in answer;
```

Fig. 3 Stochastic Matrix Construction Algorithm

4. EXPERIMENT

We run a series of simulation experiments to test our model over different network parameters and topological structures. Table 1 defines the input parameters.

We use three network types: the scale-free network is one whose degree distribution follows a power law [3]; the Erdos-Renyi random network is obtained by starting with a set of n entities and adding edges between them at random [6]; the small world network is one in which entities represent people and edges connect people who know each other, generating the small world phenomenon of strangers being linked by a mutual acquaintance [21]. In real world, many empirically observed networks appear to be scale-free, including the World Wide Web, the Internet, citation networks, and some social networks [3]. Therefore, sections 4.1-4.3 use the scale-free networks and sections

4.4 compare the results in three network topologies. All results are the average of 100 runs.

Table 1 Input Parameters of the Simulation

Input	Value	Definition
Nodesize	100, 200, 300	The total # of nodes in the network
Threshold	T=0.9, 0.8, 0.7, 0.6, 0.5, 0.4	T is the spreading threshold that is Uniform Distribution $T \in U(0,1)$
Selected-number	1, 2, 3, …, Nodesize/10	The initial selected # of nodes for targeting in cascade
Network type	Scale-free network, Erdos-Renyi random network, Small world network	Different network topologies
Density	0.01, 0.02, 0.025, 0.03, 0.035, 0.04, 0.05, 0.06	$Density = edgenum / C^2_{node}$

4.1 Dynamics vs. Node-Size and Density

In this section, we exam the network dynamics by varying the node-size and density. The total number of experiments we run is the total combinations of the values listed in Table 1:

The total number of experiments = (3 nodesize × 6 threshold × nodesize/10 selected-number × 8 density).

For each experiment, we run the simulation by using the Simulation Algorithm until all nodes in the network become active. We present the results as landscapes in Fig. 4. In Fig. 4(1), we can clearly see four landscapes, representing the spread speed of the network with four densities. The result shows that the innovation spreads faster in the network with higher density. We can see the same trend in Fig. 4(2) and Fig. 4(3) but with faster speed, especially for the network with the top three densities, they spread so fast with the similar speed that their landscapes overlap each other and looks like only one.

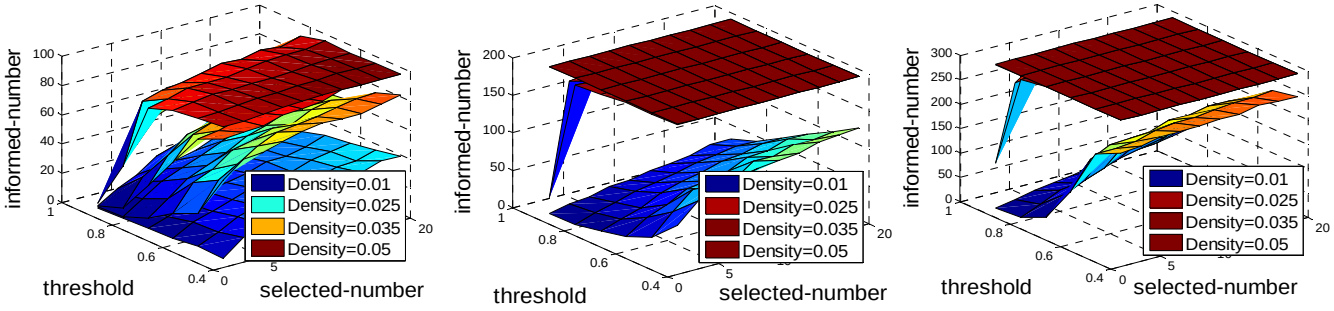
We make three conclusions based on these experiments. First, iterative diffusion in scale-free networks depends on all these factors including selected-number, threshold, and density. Second, if all other conditions are the same, the higher network density, the faster spread. In highly connected graphs, the innovation spreads easily because when a node is informed, it is likely that among its neighbors, there is someone that decides to adopt (or be informed) as well and the diffusion continues spreading. In the independent cascade model, Watts [22] observes the same phenomenon in the spreading of disease by studying real data. Thus, the iterative cascade model shares the similar characteristic with the independent cascade model in this regard. Three, if all other conditions are the same, the network with larger nodesize spread faster. The explanation is the same as the above.

4.2 Dynamics vs. Initial Selected Number

In this section, we test the final informed number by varying the initial selected-number. For each experiment, the iteration step is set to three. Fig. 5 shows that the final informed number increases with the increasing of initial selected number in the social network with different densities. Obviously, the decreasing of diffusion threshold can help to improve the final number of the informed persons, that is, the innovation can reach more people. In literature, diffusion threshold represents the *spreading tolerance* between a pair of persons. Based on our model, if the attraction of the innovation from one's friend is above his/her tolerance, he/she will be active and would like to adopt and spread the innovation too.

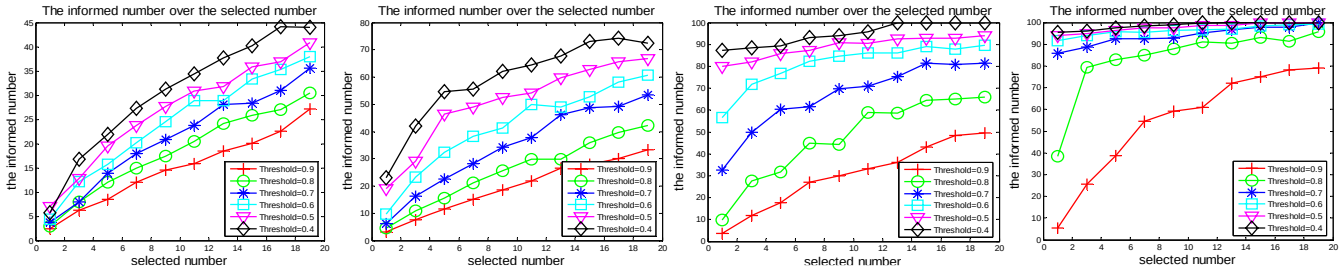
The study of diffusion threshold is important in analyzing real world problems. For example in marketing domain, the effect of mass marketing is to reduce the spreading tolerance of all the persons in the social network [23]. Previous study shows that first using mass marketing and then using viral advertising can utilize the least initial selected numbers to attain the most advertising performance, which in turn can attain the optimal performance and save money at the same time. Our experiment also reflects this idea. Let's compare the curve of threshold=0.9 and threshold=0.8 in Fig. 5(4). In the curve of threshold=0.9, the initially selected-number is three and the final informed-number is 25; while in the curve of threshold=0.8, the initially selected-number is still three but the final informed-number is 80, a great increase to 25. Also, by qualitative analysis to the curve of threshold=0.8, after mass marketing, from the point of selected-number=3, the increasing of initial selected number of persons cannot impact the final informed number more significantly than the process before selected-number=3. Based on this observation, the policy makers can define their marketing policy like this: first use mass marketing and this will generates three selected people, and then target these three persons for viral advertising and this can lead to wide adoption to their product.

The general conclusion from the above experiments is that, increasing the initial informed number can increase the final informed number. Moreover, the experiment results can provide the policy makers a concrete base for decision-making, such as the optimal point of combining mass marketing and viral advertising, and the most suitable number of initial targeting persons for viral advertising, in the above example.



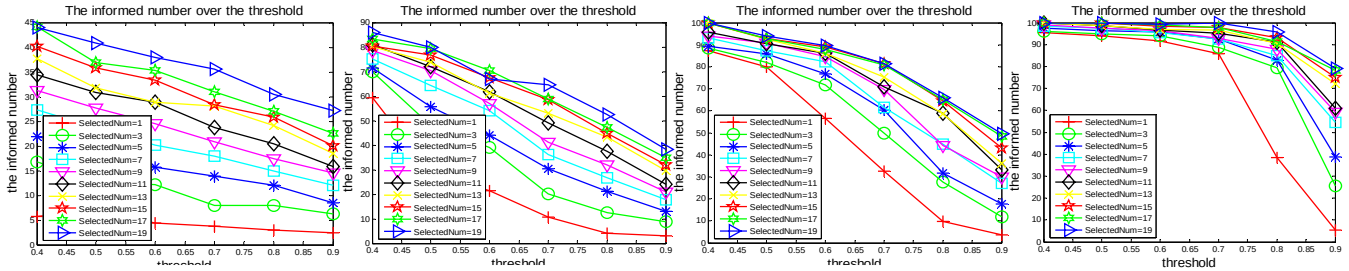
(1) Nodesize=100, type=scale-free (2) Nodesize=200, type=scale-free (3) Nodesize=300, type=scale-free

Fig. 4 Dynamics vs. Nodesize and Density



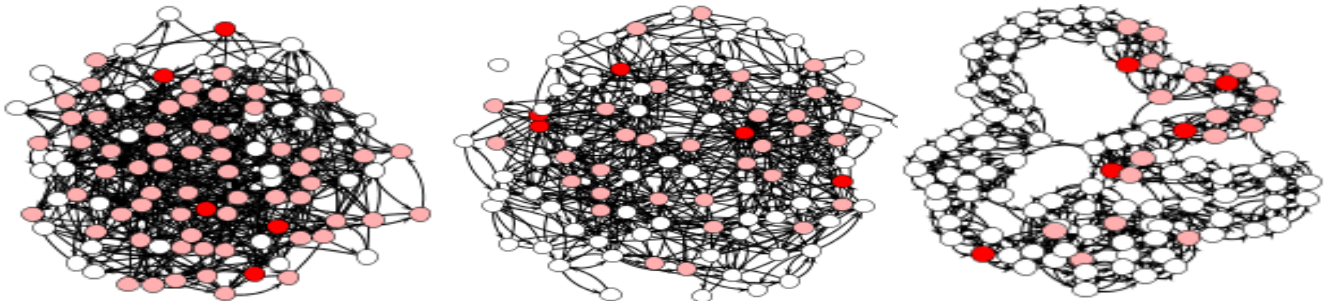
(1) Nodesize=100, density=0.005 (2) Nodesize=100, density=0.00' (3) Nodesize=100, density=0.01 (4) Nodesize=100, density=0.025

Fig. 5 The Informed Number over the Selected Number



(1) Nodesize=100, density=0.00 (2) Nodesize=100, density=0.007 (3) Nodesize=100, density=0.01 (4) Nodesize=100, density=0.025

Fig. 6 The Informed Number over the Threshold



(1) Scale-free Network (2) Random Network (3) Small-world Network

Fig. 7 Viral advertising in different network topological type (nodesize=100, density=0.005, threshold=0.9, selected-number=5. The hollow cycles represent the nodes that have not be informed; the pink circles are the nodes that have been informed finally; the red circles are the initial selected number of persons for targeting.

4.3. Dynamics vs. Diffusion Threshold

In this section, we test the final informed number by varying the diffusion threshold. The iteration step is still set to three. Fig. 6 is the results and we make two conclusions to it. First the final informed numbers increase with the decreasing of diffusion thresholds in the social networks with different densities. Second, the reducing of diffusion threshold affects advertising performance more significantly in the social network with higher density. It is shown that in the social network with density=0.005, 0.0075, and 0.01, the curves have larger slope than the case that density=0.025.

The implication of the above conclusions is domain dependent in general. Here let's still use the marketing domain as an example. These observations imply that mass marketing is more necessary for the social network with higher density. This can be rationally explained that there is less number of links in the social network with less density, so viral advertisement is harder, although mass marketing can reduce the *spreading tolerance* and promote the viral advertising. Furthermore, let's look back to Fig. 3. When density is the same and nodesize is changed (see the landscapes of density=0.01 and nodesize=100, 200, 300 respectively), the social network with higher nodesize can be benefitted more from mass marketing because the slope along the axis of threshold is larger. Therefore, the social networks with more nodes or higher density depend more on mass marketing during viral advertising.

The above analysis generates another question about how to select the suitable diffusion threshold by mass marketing in order to control spreading tolerance in viral advertising. There is an interesting observation when we focus on Fig. 6(4). The majority of curves decrease gently when the threshold increases, but after a threshold point, all curves decrease significantly. This point is threshold=0.7. This point is a tipping point that after it the advertisement will spread to a large number of people, and even everyone in the social network. Gladwell defined a tipping point as a sociological term, "the moment of critical mass, the threshold, the boiling point" [10]. To sum up, in the marketing domain, it is necessary to combine mass marketing with viral advertising, especially in the social network with high density and large node size. Discovering the tipping point by simulation is helpful on finding optimal point of mass marketing that can indeed promote viral advertising to a higher extent.

4.4. Dynamics vs. Network Topological Types

In previous sections, we have used the scale-free network to run the experiments. In this section, we compare the network dynamics in three topological structures: the scale-free network, the Erdos-Renyi random network, and the Small world network. We generate these three network structures by using ORA (Organization Risk Analyzer) that is a dynamic network analysis tool [6]. The iteration step is still set to three.

As shown in Fig. 7, relationships between the informed number and initial selected number are observed. At the same level of nodesize, density, threshold, and selected-number, the scale-free network has the best performance of spreading innovation, the Erdos-Renyi random network ranked in the middle, and the small-world network is the worst. This can be explained by the fundamental topology of these three networks. In the scale-free network, nodes share the links more even, while in the small-world network persons have better connections with their neighbors than with other persons, and there exists more cliques [20] than the other two networks. The spreading of innovation between cliques is hard due to the lack of spreading ways, so small-world network is harder to spread advertisement to all the persons than the other two networks.

5. SIMULATING A REAL-WORLD SOCIAL NETWORK IN VIRAL ADVERTISING

Finally, we apply our model to a real-world social network in viral advertising. We use datasets from Xiaonei.com which is the dominant college social network service in China (it is called as the Chinese version of Facebook). Users can access http://m.xiaonei.com/ using any mobile facilities such as IPhone, WiFi, or access http://xiaonei.com by traditional internet service. There are several attributes associated with each user, such as gender, age, hometown, online time, school, department etc. We want to answer the question that between a pair of users, which attribute will most affect them to buy a similar product?

We select the HUST (Huazhong University of Science and Technology) college network where there were 44,419 nodes and 803,987 links at the beginning of January, 2008 (data1); at the end of February, 2008, there were 47,546 nodes and 876,983 links (data2). Data1 and data2 both have the power-law degree distribution, as shown in Fig. 8.

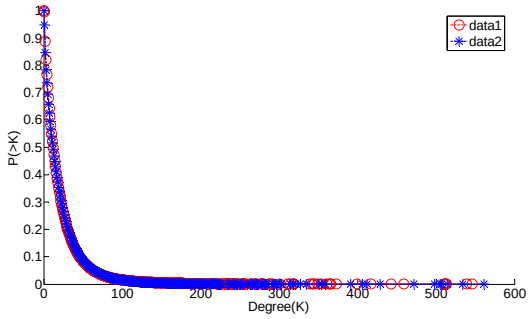


Fig. 8 Degree Distribution of Xiaonei.com

In addition, we randomly chose a subset of 603 persons (data3). First, we obtained a small fraction of the user IDs randomly; second, we used these user ids to do a snowball sampling by breadth-first search (BFS) that can well conserve the power-law nature in the degree distribution [1]. Thus, these three networks are scale-free networks.

The measurements of these three datasets are illustrated in Table 2.

Table 2 Measurements of Social networks

Measurement	Data1	Data2	Data3
Average Distance (Watts and Strogatz 1998)	4.152	4.163	5.761
Clustering Coefficient (Watts and Strogatz 1998)	0.1986	0.1874	0.0995
Density (Wasserman and Faust 1994)	0.0004127	0.0004261	0.00175
Diameter (Wasserman and Faust 1994)	44419	47546	603

A user profile sample of data3 is listed in Table 3. Person1 and Person2 are friends, and also join the same interest group, thus they have the same interest and is likely to buy the product of the same category. For the attribute "gender", the pair male-female will be calculated one time in its frequency. We calculate the frequency of all other attribute pair combinations and stored them into a MySQL database. We made an assumption that the individuals with the same interest (in the same interest groups) will buy the products in the same category. Thus, the pair of persons whose attribute pairs have the highest frequency are most likely to purchase the product of the same category.

Table 3 A User Profile Sample in HUST Social Network

Person ID	Name	Friend ID	**Out Degree**	**Gender**
100065362	Person 1	226160629 45526631	2	Male
226160629	Person 2	100065362 188274336	2	Female
Home Town	**Birth Date**	**Zodiac**	Interest Group ID	**Online Times**
Place 1	985-10-5	Scorpio	27482 43864	26
Place 2	980-3-28	Aries	27482 68306	144
Score	**Department**	**School Year**	**Last Online Time**	
103	CS	2007	2007-10-30 19:06	
450	ME	2005	2007-12-30 19:06	

In the previous Section 2, it was denoted that $W(e_{v,u})$ is the *spreading probability* that the node *v* will affect the node *u* to become informed. In this experiment, we make an assumption that $W(e_{v,u})$ is the similarity between a pair of individuals who have a probability to be interested in and buy similar products. For every pair of individuals, we calculate the weight according to the frequency:

$$W(e_{j,i}) = \frac{1}{P}\sum_{p=1}^{P} p(a_{pi}, a_{pj}), p = 1,2,...P \quad (12)$$

$$p(a_{pi}, a_{pj}) = \frac{f(a_{pi}, a_{pj})}{\sum_{n=1}^{N} f(a_{pi}, a_{pj}) + \sum_{m=1}^{M} f(a_{pm}, a_{pj})}, a_{pm} \in A(a_{pi}), a_{pm} \in A(a_{pj}) \quad (13)$$

where *P* is the total number of attributes in a personal profile and equals 10 (use bold font to represent the selected attributes in 3). $A(a_{pi})$ and $A(a_{pj})$ represent all the attributes that have been calculated as the pairs of a_{pi} and a_{pj} respectively, and *M, N* is the total number of $A(a_{pi})$ and $A(a_{pj})$, respectively.

In this experiment, the diffusion thresholds represent different mass marketing policies during viral advertising. All results are the average of 100 runs. In Fig. 7, we can see that the final informed number decreases quickly along with the increment of the threshold (the left figure), and the final informed number increases slowly along with the increment of the initial selected-number (the right figure). This is because the density of this social network is low and many nodes don't connect to each other, so there are no enough ways for spreading advertisements to all the persons. Also, in the right figure of Fig. 7, the decreasing diffusion threshold from 0.8 to 0.7 can increase the performance of final informed number larger than the decreasing diffusion threshold from 0.9 to 0.8. The pair of curves (threshold=0.9, threshold=0.8) is close to each other, and the same phenomena happened in the pairs of (threshold=0.7, threshold=0.6) and (threshold=0.5, threshold=0.5). Thus, decreasing the diffusion threshold 0.2-like decreasing from 0.9 to 0.7-is significant in this case. In Fig. 8(1), initial 1 person can spread advertisements to 307 persons, but initial 16 persons can only spread advertisement to 58 more persons (total 365 persons). This is because, as shown in Fig. 8(2), there are a lot of cliques that don't connect with each other. So increasing initial selected number cannot significantly affect the final informed number that represents the performance of spreading advertisements.

Therefore, the best solution in this case is to combine mass marketing and viral advertising: firstly invest money to reduce diffusion threshold by mass marketing; then use approximately maximization algorithm to choose 1 person as the initial target for viral advertising, and save money to invest more in mass marketing. In mass marketing, reducing the diffusion threshold one more unit than the usual policy of mass marketing-reducing the diffusion threshold one unit every time-would be helpful to increase more performance of spreading advertisements.

6. CONCLUSION

We have modeled the iterative cascade model and conducted an extensive empirical study to it by using both simulations and a real-world social network in viral advertising. These simulations have showed its power to observe and explain the emergent phenomena in the macro level when changing parameters in the micro level. One limitation of our model is its highly intractability for most graphs because it operates on a vector that is the size of the

powerset of *V*, and furthermore it must test the powerset of the target set *A* of each state *S* during every iteration of the main loop. In the future, we will simplify the model by further analyzing its own features. For example, the computation can be decreased by utilizing the fact that the low triangle of the transition matrix in Fig 2 is 0. Also we will introduce some approximation method.

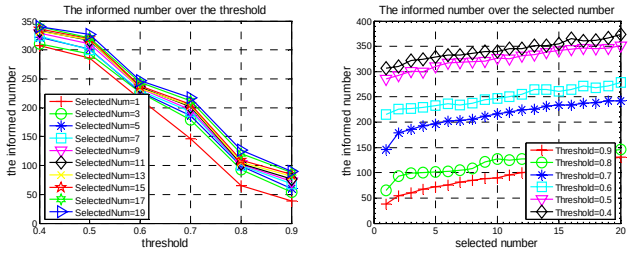
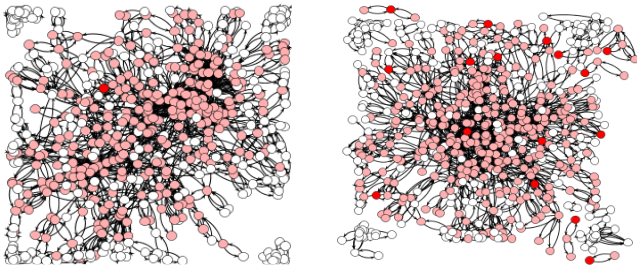


Fig. 7 Performance of the Real-World Viral Advertising



(1) Selected-number=1 (2) Selected-number=16

Fig. 8 Performance over Initial Selected-number (threshold=0.4, nodesize=603)

7. ACKNOWLEDGEMENT

We thank Stephen Foster for the initial history-dependent diffusion model when he was a NSF REU student of Yu Zhang.

REFERENCES

[1] Ahn Y, Han S, Kwak H, *Analysis of Topological Characteristics of Huge Online Social Networking Services*, International World Wide Web Conference, 2007.

[2] Albert R, Jeong H and Barabasi A, *Error and Attack Tolerance of Complex Networks*, Nature, 406: 378-382, 2000.

[3] Albert R, Barabási AL, *Statistical Mechanics of Complex Networks*, Reviews of Modern Physics, 74(1): 47-97, 2002.

[4] Bass F, *A New Product Growth Model for Consumer Durables*, Management Science, 15: 215-227, 1969.

[5] Brown J and Reinegen P, *Social Ties and Word-of-Mouth Referral Behavior*, Journal of Consumer Research, 14(3): 350-362, 1987.

[6] Carley KM, Diesner J, Reminga J, Tsvetovat M, *Toward an Interoperable Dynamic Network Analysis Toolkit*, Decision Support Systems, Special Issue on Cyberinfrastructure for Homeland Security: Advances in Information Sharing, Data Mining, and Collaboration Systems, 43(4): 1324-1347, 2007.

[7] Cowan R and Jonard N, *Network Structure and the Diffusion of Knowledge*, Journal of Economic Dynamics and Control, 28(8): 1557-1575, 2004.

[8] Culotta A, *Maximizing Cascades in Social Networks*, University of Massachusetts, Amherst, MA, Technical Report, 2003.

[9] Domingos P and Richardson M, *Mining the Network Value of Customers*, The 7th International Conference on Knowledge Discovery and Data Mining, pp, 57-66, 2001.

[10] Gladwell M, *The Tipping Point: How Little Things Can Make a Big Difference*, Little Brown, LONDON, UK, 2000.

[11] Goldenberg J, Libai B, Muller E, Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth, Marketing Letters, 3(12): 211-223, 2001.

[12] Granovetter M, *Threshold Models of Collective Behavior*, American Journal of Sociology 83(6):1420-1443, 1978.

[13] Jackson MO and Yariv L, *Diffusion on Social Networks*, Public Economics, 16(1): 3-16, 2005.

[14] Kempe D, Kleinberg J, and Tardos E, *Maximizing the Spread of Influence through a Social Network*, The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137-146, 2003.

[15] Leskovec J, Krause A and Guestrin C, *Cost-effective Outbreak Detection in Networks*, KDD 2007, San Jose, California, USA, 2007.

[16] Morris S, *Contagion*, The Review of Economic Studies, 67(1): 57-78, 2000.

[17] Richardson M and Domingos P, *Mining Knowledge-Sharing Sites for Viral Marketing*, in Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining, pp, 61-70, 2002.

[18] Rogers EM, *Diffusion of Innovations*, Free Press of Glencoe, Macmillan Company, 1962.

[19] Rolfe M, *Social Networks and Threshold Models of Collective Behavior*, Preprint, University of Chicago, 2004.

[20] Wasserman S, Faust K, *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*, Cambridge University Press, Cambridge, 1994.

[21] Watts DJ, *Small Worlds: The Dynamics of Networks Between Order and Randomness*, Princeton Press, 1999.

[22] Watts DJ, *A Simple Model of Global Cascades in Random Networks*, in Proceedings of the National Academy of Sciences, 99(9): 5766-5771, 2002.

[23] Watts, DJ and Peretti J, *Viral Marketing for the Real World*, Harvard Business Review. May 1, 2007.

Yu Zhang received her Ph.D. in Computer Science at Texas A&M University in 2005. She is currently an assistant professor in the Department of Computer Science at Trinity University and the director of the Laboratory for Distributed Intelligent Agent Systems. Dr. Zhang's research falls within Agent-Based Modeling and Simulation. Her research is currently supported by three NSF grants. She is in the editorial board for five journals including The Transaction of Simulation and IEEE Transactions on Systems, Man, and Cybernetics (SMC), Part C: Applications and Reviews, and in program committee for over 10 conferences and technical groups, such as Autonomous Agents and Multi-Agent Systems. She has regularly reviewed proposals for federal agencies. She is in the organizing committee of IEEE SMC 2009 and SCS SpringSim 2009. She is Program Co-Chair of Agent Directed Simulation of SCS SpringSim 2010. She is Program Chair of IEEE Women In Engineering Central Texas Chapter (CTC), Program Chair of IEEE SMC Society CTC, and Vice Chair of IEEE Computer Society CTC. She received 2008 Trinity Distinguish Junior Faculty Award, 2007 IEEE CTC Service Award, and Best Paper Award of 2005 International Conference on Knowledge-Based & Intelligent Information & Engineering Systems.

Jiang Wu is a PhD student in Huazhong University of Science and Technology, P.R.China. He received his M.S. from Tsinghua University, China. He has rich work experiences in software development and project management. He is combining complex practical issues in management filed with computer simulation. His interests focus on developing agent-based model to study the issues from information system research such as E-government, E-commerce, information technology diffusion and adoption etc.

Bin Hu is a professor of Management College and Director of Institute of Intelligent Management and Complex System in Huazhong University of Science and Technology, P.R.China. He has a B.S. and M.S. in Wuhan University of Technology, P.R.China (1989 and 1994). He has a Ph.D. in Management Science and Engineering from Huazhong University of Science and Technology, P.R. China (1999). His research interests include Management System Simulation, Decision Support System and Artificial Intelligence. His research papers have been published in many Chinese top journals and international conference proceedings.