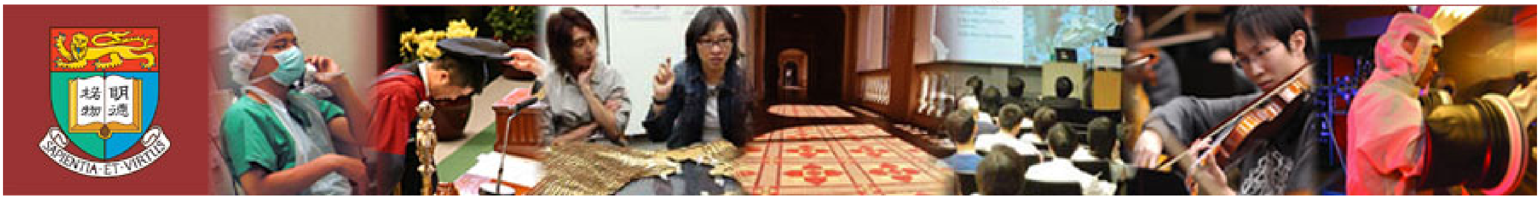


2013 IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT)

An Analytical Model for the Propagation of Social Influence

Xiaoguang Fan, Guolin Niu, Victor O.K. Li
Department of Electrical and Electronic Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong, China
Email: {xgfan, glniu, vli} @eee.hku.hk

***Abstract*—Studying the propagation of social influence is critical in the analysis of online social networks. While most existing work focuses on the expected number of users influenced, the detailed probability distribution of users influenced is also significant. However, determining the probability distribution of the final influence propagation state is difficult. Monte-Carlo simulations may be used, but are computationally expensive. In this paper, we develop an analytical model for the influence propagation process in online social networks based on discrete-time Markov chains, and deduce a closed-form equation for the n-step transition probability matrix. We show that given any initial state, the probability distribution of the final influence propagation state may be easily obtained from a matrix product. This provides a powerful tool to further understand social influence propagation.**

Keywords*—Social Network Intelligence, Social Influence Propagation, Analytical Model, Markov Chain.

I. Introduction

Social network is a complex network, connected by social relationships. Social network intelligence are widely applied in building web intelligence [1]. With the rapid development of online social networks and social media such as Twitter, Facebook, and Google+, large-scale, instantaneous information dissemination becomes possible [2]. This offers the potential for a surge of innovations and opportunities in web advertising and marketing intelligence. Various applications have been deployed, including viral marketing [3], word-of-mouth recommendation [4], and social search [5], in the hope that certain population could be targeted accurately through the diffusion process in social networks. Hence understanding the real social influence [6] and the propagation dynamics is essential as it may help inform corresponding applications and further enhance the development of web intelligence.

Motivated by the diffusion-based applications, different diffusion schemes have been proposed. Two representative models are the Threshold Model [7] and the Cascade Model [8]. The Threshold Model comes from the social science studies, and a node's tendency to become active depends on the node-specific threshold. The Cascade Model is derived from interacting particle systems studies, and each active node is given a single chance to activate its inactive neighbors. In both cases, the cascading behavior is progressive, i.e. nodes can switch from inactive to active but not in the opposite direction. Kempe et al. [9] propose a broader framework that simultaneously generalizes both models where the probability that an inactive node becomes active increases monotonically as the number of its active neighbors increases. Another research direction is to study the spread of influence where explicit network topology is unknown. Yang and Leskovec [10] propose a Linear Influence Model (LIM) to depict the global influence through an implicit network. However, LIM is empirical rather than analytical, and we cannot get an accurate description of the process evolution over time.

Given certain stochastic diffusion models, how can we calculate the expected influence spread? In other words, given a set of initial active nodes, what is the probability distribution of the final influence propagation state? For the Cascade Model and the Threshold Model, the evaluation could only be achieved through Monte-Carlo simulation techniques. Although submodularity properties [11] of the diffusion process have been extensively studied and a performance bound of $1-\frac{1}{e}$ (e is the base of natural logarithm, and it is approximately equal to 2.71828) can be guaranteed for a greedy algorithm, only the final expected influenced number but not the accurate influence propagation state distributions can be obtained [9]. A large portion of existing studies focus either on exploring more efficient algorithms to provide a good estimation for the expected number of influenced nodes [12] [13], or on developing new diffusion models [14] [15] to directly predict the influence spread based on real propagation traces. To summarize, none of the current performance evaluation approaches for the social influence propagation provides accurate probability distribution of the final influence propagation state. Thus proposing an appropriate diffusion scheme and a comprehensive performance evaluation framework is vital for designing efficient marketing strategies.

To achieve this goal, we borrow certain concepts from other disciplines. In marketing and management science, the most basic and widely-studied model is the Bass diffusion model [16], which quantitatively describes how new products get adopted as an interaction between existing and potential users. The imitation coefficient is used to describe the interaction among existing customers. In epidemiological studies [17], the disease or virus propagation process are described using certain parameters like contact rate and recovery rate. Existing applications of epidemiology models on Online Social Network (OSN) studies produce useful quantitative results [18] [19]. Inspired by the similarities in the patterns of product adoption diffusion, epidemic propagation and the information cascading process in OSN, we propose a new Markov chain based Reinforced Cascade Diffusion (MRCD) model, where a diffusion parameter p is used to represent the activation

978-1-4799-2902-3/13 $31.00 © 2013 IEEE
DOI 10.1109/WI-IAT.2013.2

probability of each node, and repeated tries to activate the same node is allowed. Specifically, we build an analytical framework for the information diffusion process based on discrete-time Markov chains [20] [21], and deduce a closed-form equation to express the n-step transition probability matrix. We show that given any initial state the probability distribution of the converged network state could be easily obtained by calculating a matrix product. With the analytical framework, we give a good estimation of the final influence propagation state.

Our contributions in this paper can be summarized as follows.

- We introduce an MRCD model and deduce a closed-form equation to express the probability distribution of the influence propagation state.
- We propose a *ConstructMatrix* algorithm to calculate the 1-step transition probability matrix.
- We evaluate our analytical model in regular graphs and study the impact of network topology on the probability distribution of the final influence propagation state.

This paper is structured as follows: The diffusion process is described in Section II, our proposed analytical model in Section III, the *ConstructMarix* algorithm in Section IV, model evaluation in Section V, implications in Section VI, and the conclusion and future work in Section VII.

II. DIFFUSION PROCESS

Considering a social network with N nodes as a directed graph $G(V,E)$ where V is the set of vertices and E is the set of edges, we define the in-degree and out-degree neighbor set of Node j as $N_j^{in} = \{i \in V : (i,j) \in E\}$ and $N_j^{out} = \{i \in V : (j,i) \in E\}$, respectively. Here edge (i,j) is directed from Node i to Node j. Each node is active or inactive. Once a node becomes active, it cannot return to being inactive. Originally all of the nodes are inactive, and a set of nodes, namely, the seed set, is made active initially. In the following discrete steps, nodes are activated by their active in-degree neighbors and in turn activate their inactive out-degree neighbors. For instance, Node i that becomes active at Step t has a probability p to successfully activate each inactive out-degree neighbor j through edge (i,j) in Step $t+1$. The probability is independent of the historical activation track. Note that the order of activation is random when there are multiple active neighbors trying to activate a common node. The process is finished when no more activation is available. We monitor the network status at each step. The process ends if there is no new activations at Step t compared with Step $t-1$.

III. DISCRETE-TIME MARKOV CHAIN MODEL

A Markov chain is a sequence of states $X_i (i = 1...n)$ which are random variables with Markovian property, described as follow:

$$\begin{aligned} Pr(X_{n+1} = x | X_n = x_n, X_{n-1} = x_{n-1}, ..., X_1 = x_1) \\ = Pr(X_{n+1} = x | X_n = x_n) \end{aligned} \tag{1}$$

It shows that the future and past states are independent given the current state. Here the set with all possible value of x_i is called the state space of the chain, and $Pr(X_{n+1} = x | X_n = x_n)$ is called the state transition probability. Suppose that the Markov chain is discrete and the state space is finite, we could build the 1-step transition probability matrix P_1.

$$P_1 = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,n} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n,1} & p_{n,2} & \cdots & p_{n,n} \end{pmatrix} \tag{2}$$

where n is the total number of states and $p_{i,j}$ is the transition probability from State i to State j in one time step.

A. Network state

Assume that the target consumer group has n nodes with ID $0,1,...,n-1$. We define the state of each node as 1 if it is active, and 0 otherwise. Thus, the whole network state may be represented as a binary sequence. For instance, a three-node network has eight states $\{0,1,2,3,4,5,6,7\}$. State 0 with binary code 000 means all the nodes in the system are inactive. State 6 with binary code 110 means Nodes 0 and 1 are active while Node 2 is inactive. In order to judge whether the state is converged in the diffusion process, we define both the stable and unstable states for each binary sequence. For the binary sequence with decimal value i, we define i as the unstable state and $\widehat{i}$ as the stable state. Thus a network with n nodes has 2^{n+1} states $\{0,1,...,2^n - 1,\widehat{0},\widehat{1},...,\widehat{2^n - 1}\}$.

B. Transition probability matrix

Next we build a $2^{n+1} \times 2^{n+1}$ 1-step transition probability matrix for the n-node network. Both row and column indices correspond to the states $\{0,1,...,2^n - 1,\widehat{0},\widehat{1},...,\widehat{2^n - 1}\}$. We have the following rules to determine the value of the elements in the matrix.

1) $p_{i,\widehat{j}} = 0 \; if \; i \neq j$
If the current network state is unstable, it may have two choices, namely, moving to other unstable states or to its corresponding stable state.
2) $p_{i,j} = 0, \; if \; i \geq j$.
Since the active node cannot become inactive, the unstable network state with a larger decimal value cannot move to the unstable state with a smaller decimal value.
3) $p_{\widehat{i},\widehat{j}} = 1, \; if \; i = j$ and $p_{\widehat{i},\widehat{j}} = 0, \; if \; i \neq j$
If the current network state is stable, it cannot move to other states but must remain in this state to the end. Thus the diffusion process will finally converge to one stable state.
4) $p_{\widehat{i},j} = 0$
A stable state cannot move to other unstable states.

According to the rules listed above, we may represent our 1-step transition probability matrix as a partitioned matrix:

$$P_1 = \begin{pmatrix} A & B \\ O & I \end{pmatrix} \tag{3}$$

where A is an upper triangular matrix according to Rule 2, B is a diagonal matrix according to Rule 1, O is a zero matrix according to Rule 4, and I is an identity matrix according to Rule 3. All the matrix blocks are $2^n \times 2^n$.

Next we calculate the n-step transition probability matrix P_n. Since $P_n = {P_1}^n$, we have

$$P_n = \begin{pmatrix} A & B \\ O & I \end{pmatrix}^n = \begin{pmatrix} A^n & (\sum\limits_{i=0}^{n-1} A^i)B \\ O & I \end{pmatrix} \tag{4}$$

where $A^0 = I$. Considering the final probability distribution of network state, we only focus on the upper-right block in the matrix, since this block includes all the probabilities from unstable states to stable states. We denote it as G and have the following deduction:

$$\begin{aligned} G &= (\sum_{i=0}^{n-1} A^i)B \\ (I-A)G &= (I-A)(\sum_{i=0}^{n-1} A^i)B \\ (I-A)G &= (I-A^n)B \\ G &= (I-A)^{-1}(I-A^n)B \end{aligned} \tag{5}$$

According to Rule 2, the diagonal elements of matrix A are all zeros, then the diagonal of the matrix $I-A$ are all ones. Since $I-A$ is still an upper triangular matrix, it is guaranteed to be a full-rank matrix and thus invertible.

C. Convergence state

Thus far we have given an expression of the n-step transition probability matrix in (5). Next we discuss the convergence issue. To analyze the convergence state of matrix G, we need to first deal with the component A^n. A is an upper triangular matrix with all the diagonal elements being zeros. According to linear algebra theory, there exists an invertible matrix U such that

$$A = U^{-1}JU \tag{6}$$

Here J is a Jordan matrix

$$J = \begin{pmatrix} J_{\lambda_1,m_1} & & & \\ & \ddots & & \\ & & J_{\lambda_i,m_j} & \\ & & & \ddots \end{pmatrix} \tag{7}$$

J is a block diagonal matrix, and it is composed of Jordan blocks J_{λ_i,m_j}, ($i = 1,2,..., j = 1,2,...$):

$$J_{\lambda_i,m_j} = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix} \tag{8}$$

The Jordan block J_{λ_i,m_j} has a size $m_j \times m_j$. It is composed of zero elements everywhere except for the diagonal, which is filled with λ_i, and for the superdiagonal, which is composed of ones. Here λ_i is the eigenvalue of matrix A.

Since A is an upper triangular matrix with all diagonal elements being zeros, it only has the eigenvalue 0. The matrix J could be simplified as follow:

$$J = \begin{pmatrix} J_{0,m_1} & & \\ & \ddots & \\ & & J_{0,m_k} \end{pmatrix} = \bigoplus_{i=1}^{k} J_{0,m_i} \tag{9}$$

where J_{0,m_i} is an $m_i \times m_i$ Nilpotent matrix:

$$J_{0,m_i} = \begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix} \tag{10}$$

The $m_i \times m_i$ Nilpotent matrix J_{0,m_i} satisfies $J^n_{0,m_i} = 0$ ($n \geq m_i$). Thus A^n could be calculated as follow:

$$\begin{aligned} A^n &= (U^{-1}JU)^n \\ &= (U^{-1}JU)(U^{-1}JU)...(U^{-1}JU) \\ &= U^{-1}J^nU \\ &= U^{-1}(\bigoplus_{i=1}^{k} J_{0,m_i})^n U \\ &= U^{-1}(\bigoplus_{i=1}^{k} (J_{0,m_i})^n) U \end{aligned} \tag{11}$$

Suppose that the maximum size among all the Jordan blocks $J_{0,m_i}(i = 0,1...,k)$ is m, when $n > m$, $J_{0,m_i} = 0$ $(i = 0, 1..., k)$. Thus $A^n = 0$. It means that when the time steps are big enough, the probability distribution of the network state represented by the matrix G converges to a stable value:

$$G = (I-A)^{-1}B \tag{12}$$

where I is an identity matrix, A and B are the upper left and upper right blocks of the 1-step transition probability matrix P_1, respectively. Thus, given the 1-step transition probability matrix P_1, we could calculate the final probability distribution of network state.

IV. 1-STEP TRANSITION PROBABILITY MATRIX CONSTRUCTION

In this section, we describe how to construct the transition matrix by developing the *ConstructMatrix* algorithm and also provide the pseudo code for *ConstrucMatrix*.

A. Algorithm structure

The structure of our algorithm is shown in Fig. 1. The input of *ConstrucMatrix* includes three parts:

- **Network topology** is defined in Section II. We represent it as an $n \times n$ matrix T. $T_{i,j} = 0$ $if(i,j) \notin E$, and $T_{i,j} = p_{i,j}$ otherwise. Here $p_{i,j}$ is the activation probability from Node i to Node j.
- **System state** is defined in Section III-A

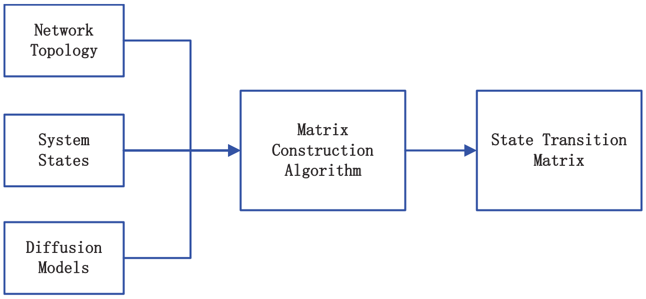


Figure 1. 1-step transition probability matrix construction process

- **Diffusion model** is defined in Section II. Note that the diffusion rules may influence the structure of the 1-step transition probability matrix P_1.

The output of *ConstrutMatrix* is the 1-step transition probability matrix P_1. Due to the properties of P_1 described in (3), we will describe how to construct the two block matrices A and B.

B. Matrix A

Firstly, we will build matrix A. Given two arbitrary network states i and j, we introduce three vectors I,J_1 and J_0 to calculate the transition probability p_{ij} from State i to State j. I, J_1, J_0 are utilized to store the active node IDs in State i, the newly added active node IDs from State i to State j, and the remaining inactive node IDs from State i to State j, respectively. For example, for a 4-node network, we assign node IDs as $\{0,1,2,3\}$. Given two system states $i = 4$ (with binary code 0100) and $j = 7$ (with binary code 0111), we can get vector $I = [1]$, $J_1 = [2,3]$, and $J_0 = [0]$. Based on the network topology represented by matrix T and the network state described by vectors I,J_1 and J_0, the transition probability can be calculated by Algorithm 1.

```
Algorithm 1: Build matrix $A$
  Input:
    $n * n$ topology matrix $T$, system state vector
    $S = [0,1,2,...,2^n - 1]$.
  Output:
    The constructed state transition matrix $A$;
    for $i = 0$ to $2^n - 1$ do
      for $j = 0$ to $2^n - 1$ do
        if $i \geq j$ OR $i \,\&\, \overline{j} == 0$ then
          $A_{ij} = 0$
        else
          calculate vector $I$, $J_1$, and $J_0$
          $A_{ij} =$ calProbability($I$,$J_1$,$J_0$)
        end if
      end for
    end for
    return $A$
```

C. Matrix B

Based on matrix A, matrix B is generated by Algorithm 3.

```
Algorithm 2: calProbability($I$,$J_1$,$J_0$)
  Input:
    Topology matrix $T$, vector $I = [i_1,i_2,...i_x,...]$,
    $J_1 = [j_{11},j_{12},...j_{1y},...]$, and $J_0 = [j_{01},j_{02},...j_{0y},...]$.
  Output:
    The final output probability from state $I$ to $J$ is $p_{I \to J}$
    $p_a = \prod_{j_{1y} \in J_1} [1 - \prod_{i_x \in I} (1 - T_{i_x,j_{1y}})]$
    $p_b = \prod_{j_{0y} \in J_0} \prod_{i_x \in I} (1 - T_{i_x,j_{0y}})$
    $p_{I \to J} = p_a \times p_b$
    return $p_{I \to J}$
```

```
Algorithm 3: Build matrix $B$
  Input:
    Matrix $A$
  Output:
    The final output stable transition matrix is $B$
    for $i = 0$ to $2^n - 1$ do
      $B_{ii} = 1 - \sum_{j=1 \sim 2^n} A_{ij}$
    end for
    return $B$
```

V. Model Evaluation

In this section, we try to apply our MRCD model on some sample graphs. First we evaluate our model in comparison with Monto Carlo method, and then we show the total probability distribution of the final influence propagation state, and finally we study the impact of network topology on this distribution.

A. Sample Graph

In order to focus on evaluating the performance of our analytical model, we hope to remove the effect of uncertainty on graph topology. Thus we choose regular graphs introduced in [22] to test our model. These graphs have the following properties:

- The graph is centro-symmetric, which means all nodes are identical in the network.
- Each node has the same fixed degree, and the graphs are identified by the degree value.

The regular graphs we used are shown in Fig. 2. Here each graph has eight nodes, and the degree values are from 2 to 7. Fixing the degree for each node is easily achieved by the following rules.

- To construct 2-degree graph, we may connect all nodes in a circle. It is shown in the first graph in Fig. 2.
- To construct $2n$-degree graph ($n = 2,3,...$), we could further connect each node with its 2-hop, 3-hop, ..., and n-hop neighbors based on the 2-degree graph. For instance, we may connect 2-hop neighbors in the 2-degree graph to build the 4-degree graph shown in Fig. 2.

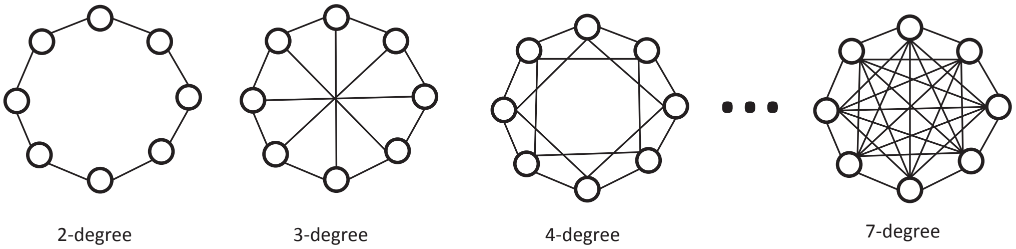


Figure 2. Regular graphs with different degree values

Table I. PERFORMANCE COMPARISON BETWEEN THE MONTE CARLO METHOD AND OUR MRCD MODEL

Settings	Sample Mean	Sample Deviation
Avg_{100}	1.100	0.0307
Avg_{1000}	1.100	0.0098
Avg_{10000}	1.100	0.0031
Avg_{MRCD}	1.100	N/A

- To construct $2n+1$-degree graph (n = 1,2,...), we could make extra link between each node and its centro-symmetric node base on the corresponding $2n$-degree graph. For instance, we can connect the centro-symmetric node pairs in the 2-degree graph to build the 3-degree graph shown in Fig. 2.

Moreover, we set an identical activation probability $p = 0.05$ for each edge in all the graphs.

B. Comparison with Monte Carlo method

We run the Monte Carlo method, as well as our analytical model, to estimate the final influence propagation, and compare the results. We take the 2-degree regular graph as an example. Since all the nodes are identical in a 2-degree graph, we choose one node as the active seed, and then utilize the Monte Carlo method and the analytical model to run the diffusion process described in Section II. The performance metric is the average number of active nodes at the end of the diffusion process.

For the Monte Carlo method, to obtain this performance metric, we consider three experiments with 100, 1000, and 10000 runs. In a run, we simulate the whole information diffusion process one time, and get the number of active nodes at the end. Then we calculate the performance metric for the three experiments, and denote them Avg_{100}, Avg_{1000} and Avg_{10000}, respectively. Each sample mean, e.g. Avg_{100}, is the result of one experiment. To determine the distribution of each sample mean, we repeat each experiment 10000 times, obtaining three distributions of sample means, as shown in Fig. 3. For our MRCD model, we utilize the *ConstrucMatrix* algorithm described in Section II to calculate the 1-step transition probability matrix, and then utilize (12) to calculate the distribution of influence propagation state. Fixing the start state at 1, the corresponding row of the matrix G is the probability distribution from State 1 to all other stable states. Through this distribution, we can easily calculate the average number of activations, and denote it Avg_{MRCD}.

Fig. 3 shows the distributions of Avg_{100}, Avg_{1000}, and Avg_{10000} calculated by the Monte Carlo method. According to the central limit theorem, the three distributions follow the Gussian distribution. Moreover, when the number of simulation runs increases, the deviation of the distribution decreases. The sample means and deviations of Avg_{100}, Avg_{1000}, and Avg_{10000}, as well as Avg_{MRCD} from our analytical model are shown in Table I. According to statistical theory, the sample mean is the unbiased estimate of the average number of active nodes. The value calculated from our analytical model is the same as the sample mean. Moreover, even for the experiment of 10000 simulation runs, there is still a deviation of 0.0031 in estimating the average number of active nodes. By contrast, our analytical model gives the accurate result without variance.

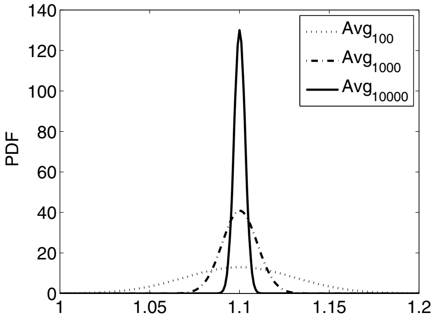
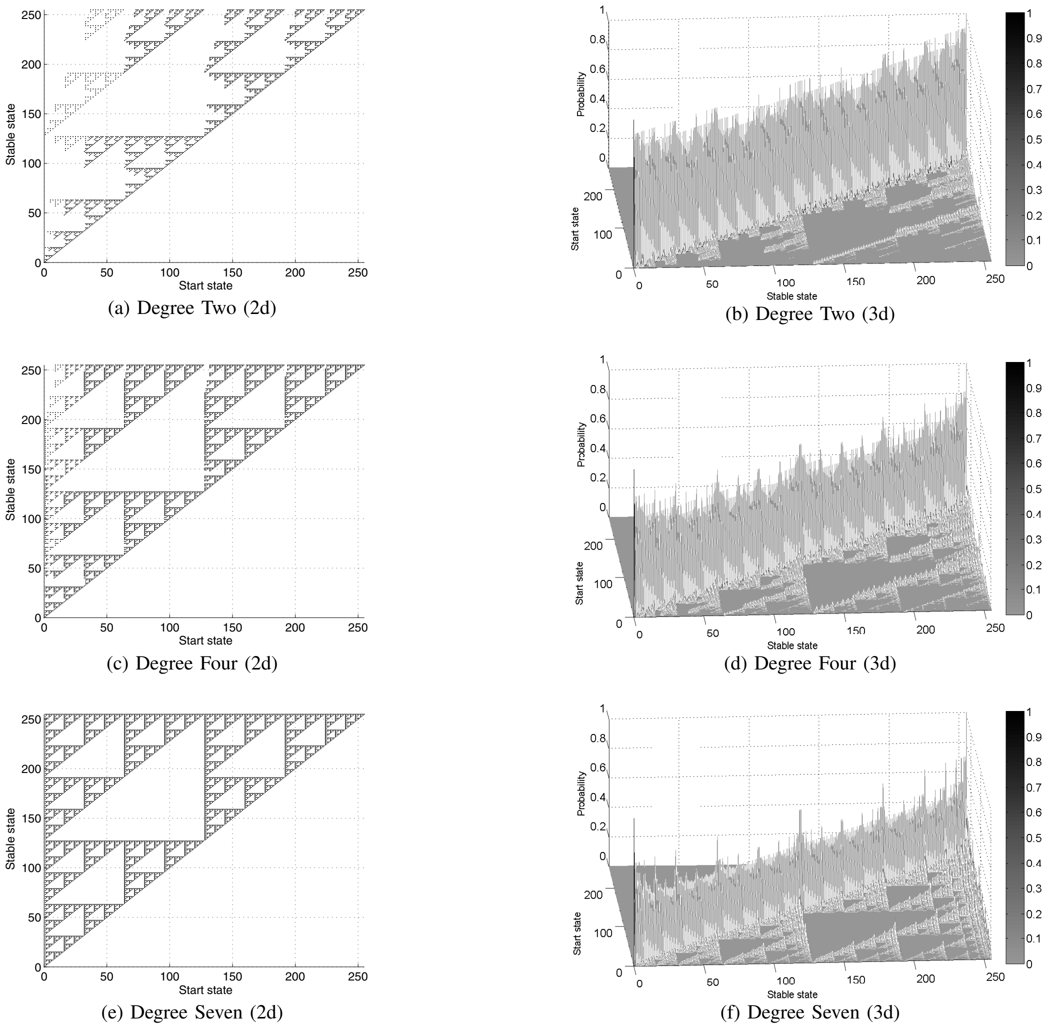


Figure 3. The distributions of Avg_{100}, Avg_{1000}, and Avg_{10000}

C. Propagation state distribution

Fig. 4 shows the probability distribution of the influence propagation state for the 2-degree, 4-degree and 7-degree regular graphs respectively[1]. Fig. 4(b), 4(d), and 4(f) are the three-dimensional (3D) graphs. It is generated by the mesh network through Matlab. The horizontal plane has two dimensions which represent the start state and final stable state, respectively, and the vertical axis represents the transition probability from the start state to the stable state. In order to better describe the distribution, we also draw the two-dimensional (2D) scatter graphs in Fig. 4(a), 4(c), and 4(e). They are the top views of the corresponding 3D graphs. The x axis is the start state, and the y axis is the stable state. Point (x, y) is blank if the transition probability from start state x to stable state y is lower than the threshold Θ. Here we assign Θ as 10^{-5}, and the value under 10^{-5} could be regarded as zero.

[1] The result of 3-degree, 5-degree, and 6-degree graphs show the same trend. We omit them due to space limitations.



(a) Degree Two (2d)

(b) Degree Two (3d)

(c) Degree Four (2d)

(d) Degree Four (3d)

(e) Degree Seven (2d)

(f) Degree Seven (3d)

Figure 4. Probability distribution of influence propagation state for 2-degree, 4-degree and 7-degree regular graphs in 2D and 3D format

Note that the state values shown on the x and y axes are the decimal values from the binary code described in Section III-A. We enlarge Fig. 4(f) to Fig. 5 in order to gain more details of the distribution. Fig. 5 represents the fully connected graph. There are several peaks as well as blank triangle areas. We can see that the points with high transition probability gather at the diagonal band, while points with low transition probability are scattered to the right half of the diagonal.

Next we turn to analyze the 2D graphs shown in Fig. 4(a) 4(c) and 4(e). Look at Fig. 4(e) first. It represents the fully connected graph. We have the following observations:

- The lower-right triangle-shape area below the curve $y = x$ is blank. This is due to Rule 2) in Section III-B. The unstable state cannot move to the unstable state with a smaller decimal value. Thus it is impossible to move to the stable state with a smaller decimal value. This big blank triangle is not determined by the network topology, but only by Rule 2).
- We can also see that there are several symmetric triangle-shape blanks in the upper-left area. This is due to the rule of our diffusion model described in Section III-B. Since the active node cannot return to being inactive, the transition from 1 to 0 is not allowed. Thus there is a fixed pattern of symmetric blank triangles in Fig. 4(e), and these blank triangles are independent of the network topology.

Then we compared the results in Fig. 4(a) 4(c) and 4(e). We can see that when degree increases, which means the density of the network rises, there are more points covered, especially in the upper-left square area $\{(x,y) : x \in [0,50], y \in [200,255]\}$. Since the start states from 0 to 50 are regarded as the states with few active nodes, and the stable states from 200 to 255 are regarded as the states with many active nodes, this shows that increasing the network density will increase the coverage of the stable states which has a greater number of active nodes, thereby enhancing the influence propagation.

For the corresponding 3D cases, by comparing the distributions in Fig. 4(b), 4(d), and 4(f), we can see that when

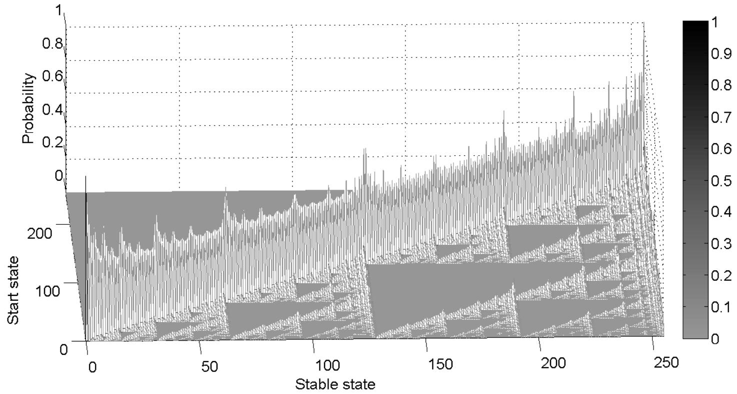


Figure 5. The probability distribution of the influence propagation state for fully connected regular graph

degree increases, the transition probabilities decrease in the diagonal band, and increases outside of the diagonal band. Since the density of the regular graph solely depends on the degree value, this shows that the rise of network density may help to balance the transition probabilities between different states, and further reduce the variance of the total distribution.

Finally we summarize the general observations:

- There are fixed patterns of triangle-shape areas with zero transition probabilities, and they are independent of the network topology.
- From the 2D graphs, when the degree increases, there are more points covered. This shows that increasing the network density will enhance the influence propagation.
- From the 3D graphs, when degree increases, the transition probabilities decrease in the diagonal band, and increases outside of the diagonal band. This shows that the rise of network density may help to reduce the variance of the probability distribution.

VI. Implications

The most significant business application on social influence propagation is social media marketing. With the limited budget on advertising, the major goal is to utilize a few influential nodes to trigger a large cascade of activations through the online social media. The influence maximization problem in online social networks has been widely studied (e.g. [9], [23], [12]). However, most of these efforts focus on increasing the expected final active set size while ignoring the explicit probability distributions. We list several implications to demonstrate the importance of exploring the probability distributions of activations.

A. Risk measure and control

While excessively emphasizing the efficiency of social media marketing in terms of low advertizing expense, rapid information dissemination and well-executed marketing strategy, most efforts ignore one significant aspect in the marketing strategy valuation. That is, the risk measurement and control mechanism. Explicit definition and measurement for the risk in social media marketing should be given, so we can further explore how to avoid them through quantitative and qualitative ways. Generally the risk in social media marketing could be regarded as the measurement uncertainty on performance.

Currently, utilizing social media in order to trigger big influence cascade is still at the theoretical studies stage, with rarely any real business applications. One reason is the lack of an accurate evaluation on the performance of social media marketing. Businesses cannot be convinced by equations and theorems. They need measurable and quantifiable results on the effectiveness of proposed strategies, such as Return On Investment (ROI). Giving a reliable measurement and estimation method on performance is the key to determining the business value of a strategy. However, currently the uncertainty of the performance is the main obstacle to social media marketing deployment.

Since information diffusion is a stochastic process, all the variables are probabilistic. The final result of a specific social media marketing strategy is a random variable with a featured probability distribution function. Although we could use the expected value to give an approximation, the variance cannot be ignored. This could be regarded as a risk in measuring the final performance. Thus, knowing explicitly the total probability distribution function will be helpful for us to do the risk control. The analytical influence model proposed in Section III gives a good estimation of the final probability distribution by using simple matrix equation, thus providing an efficient tool in controlling the risk of performance measurement.

B. The impact of topology on activation distributions

The distribution of activations would help us further define the efficient social network. Here the efficient social network is regarded as the social network in which fluent information diffusion is available, and influence propagation is easily triggered. Since the efficiency of social network mainly depends on the network topology and the diffusion mechanism (activation probability), with our proposed MRCD model, we could easily analyze the impact of topology on the final activation distribution. In this paper, as shown in Section V, we do a preliminary study on regular graphs. In the future, we may extend the problem to more general graphs.

C. Maximizing influence in heterogeneous social network

The traditional influence maximization problem solely focuses on maximizing the expected number of active nodes based on a fixed number of active seeds at the beginning of the activation process. The underlying assumption is that the social network is a homogeneous network, in which nodes are indistinguishable in activation costs and purchasing abilities. If the factor of individual difference or group difference is taken into consideration, the final active set size may not be an accurate metric to evaluate the performance of social marketing strategies. In other words, the distribution of activations need to be investigated in detail. Moreover, in some cases, we hope to influence some favorable nodes while avoiding to activate unfavorable nodes. Sometimes the activation of favorable nodes brings positive effect, while activation of unfavorable nodes may cause negative impact. In order to realize precision marketing, we also need to find out the distribution of activations.

VII. Conlusion and Future work

In this paper, we concentrate on modeling the discrete-time diffusion process. We build an analytical model for the influence propagation process based on discrete-time Markov chains, and deduce a closed-form equation to express the n-step transition probability matrix. In the future, we shall find efficient ways to reduce the computational cost of the matrix in our analytical model. For instance, we may conduct state classification and aggregation to reduce the size of the state space. Moreover, based on the analytical results, we could get the final probability distribution of the influence propagation state. This may help us further understand the influence propagation process. For instance, based on the distribution gained, we could measure and control the risk of the social media strategy. In addition, we may study the effect of social network topology on the propagation of influence.

References

[1] T. Nishida, "Social intelligence design and human computing," in *Artifical Intelligence for Human Computing*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007, vol. 4451, pp. 190–214. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-72348-6_10

[2] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12. New York, NY, USA: ACM, 2012, pp. 519–528. [Online]. Available: http://doi.acm.org/10.1145/2187836.2187907

[3] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Trans. Web*, vol. 1, no. 1, May 2007. [Online]. Available: http://doi.acm.org/10.1145/1232722.1232727

[4] J. Huang, X.-Q. Cheng, H.-W. Shen, T. Zhou, and X. Jin, "Exploring social influence via posterior effect of word-of-mouth recommendations," in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, ser. WSDM '12. New York, NY, USA: ACM, 2012, pp. 573–582. [Online]. Available: http://doi.acm.org/10.1145/2124295.2124365

[5] D. Carmel, N. Zwerdling, I. Guy, S. Ofek-Koifman, N. Har'el, I. Ronen, E. Uziel, S. Yogev, and S. Chernov, "Personalized social search based on the user's social network," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ser. CIKM '09. New York, NY, USA: ACM, 2009, pp. 1227–1236. [Online]. Available: http://doi.acm.org/10.1145/1645953.1646109

[6] S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proceedings of the National Academy of Sciences*, vol. 106, no. 51, pp. 21 544–21 549, 2009. [Online]. Available: http://www.pnas.org/content/106/51/21544.abstract

[7] M. Granovetter and R. Soong, "Threshold models of diffusion and collective behavior," *The Journal of Mathematical Sociology*, vol. 9, no. 3, pp. 165–179, 1983. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/0022250X.1983.9989941

[8] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing Letters*, vol. 12, pp. 211–223, 2001. [Online]. Available: http://dx.doi.org/10.1023/A%3A1011122126881

[9] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '03. New York, NY, USA: ACM, 2003, pp. 137–146.

[10] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," in *Proceedings of the 2010 IEEE International Conference on Data Mining*, ser. ICDM '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 599–608. [Online]. Available: http://dx.doi.org/10.1109/ICDM.2010.22

[11] M. Fisher, G. Nemhauser, and L. Wolsey, "An analysis of approximations for maximizing submodular set functions," in *Mathematical Programming Studies*. Springer Berlin Heidelberg, 1978, vol. 8, pp. 73–87. [Online]. Available: http://dx.doi.org/10.1007/BFb0121195

[12] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '10. New York, NY, USA: ACM, 2010, pp. 1029–1038. [Online]. Available: http://doi.acm.org/10.1145/1835804.1835934

[13] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *Proceedings of the 2010 IEEE International Conference on Data Mining*, ser. ICDM '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 88–97. [Online]. Available: http://dx.doi.org/10.1109/ICDM.2010.118

[14] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "A data-based approach to social influence maximization," *Proc. VLDB Endow.*, vol. 5, no. 1, pp. 73–84, Sep. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=2047485.2047492

[15] X. Song, Y. Chi, K. Hino, and B. L. Tseng, "Information flow modeling based on diffusion rate for prediction and ranking," in *Proceedings of the 16th International Conference on World Wide Web*, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 191–200. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242599

[16] F. M. BASS, "A new product growth for model consumer durables," *MANAGEMENT SCIENCE*, vol. 15, no. 5, 1969.

[17] F. Brauer, P. Van den Driessche, and J. Wu, *Lecture Notes in Mathematical Epidemiology*. Springer, 2008.

[18] L. M. Bettencourt, A. Cintrn-Arias, D. I. Kaiser, and C. Castillo-Chvez, "The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models," *Physica A: Statistical Mechanics and its Applications*, vol. 364, no. 0, pp. 513 – 536, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378437105010113

[19] F. J. Santonja, A. C. Tarazona, and R. J. Villanueva, "A mathematical model of the pressure of an extreme ideology on a society," *Comput. Math. Appl.*, vol. 56, no. 3, pp. 836–846, Aug. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.camwa.2008.01.001

[20] J. R. Norris, *Markov chains*. Cambridge university press, 1998, no. 2008.

[21] R. R. Sarukkai, "Link prediction and path analysis using markov chains," *Comput. Netw.*, vol. 33, no. 1-6, pp. 377–386, Jun. 2000. [Online]. Available: http://dx.doi.org/10.1016/S1389-1286(00)00044-X

[22] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-worldnetworks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[23] X. Fan and V.-K. Li, "The probabilistic maximum coverage problem in social networks," in *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, 2011, pp. 1–5.