“Two is a Crowd” - Optimal Trend Adoption in Social Networks

Lilin Zhang, Peter Marbach
Department of Computer Science
University of Toronto
{*llzhang, marbach*}@cs.toronto.edu

Abstract—**In this paper, we study how an individual in a social network should decide whether or not to adopt a trend, based on how many people in his/her neighborhood in the social network adopted the trend. In particular we are interested in the question in what adoption policy leads to an optimal trend-adoption in the sense that an individual only adopts a trend if the majority of his/her social network will do so. We use a decision process on a Erdös-Rényi random graph model to model and study this situation. Using this model, we obtain the result that the optimal policy for an individual is to adopt a trend if two people in their neighborhood did. Interestingly, this result/behavior was experimentally observed in real social networks. We hope that this work will help towards building applications that is able to automatically push relevant content to users in online social networks.**

I. Introduction

Online Social Networks have experienced an explosive growth over the recent a few years, and that growth has accompanied abundant interest in the study of the spread of trends in social networks. In this paper, we study how an individual in a social network should decide on whether or not to adopt a trend. We use the term “trend” quite broadly where a trend could for example be a fashion trend, a new technology, an idea, or a behavior.

Two paradigms for modeling trend adoption in social networks have been proposed in the literature: the threshold model and the cascade model. Both models are based on the natural assumption that individuals tend to be influenced by their acquaintances with whom they communicate. Yet the two are intrinsically modeling two different types of influence in social networks. In the threshold model, each individual observes the number of adoptions among his friends, and only adopts the trend if the number of adoptions among his/her friends exceeds a particular threshold value. On the other hand, in the cascade model, each individual can be convinced to adopt the trend through interactions with friends in the social network who have already adopted the trend.

However, two recent studies by Backstorm el. [4] on friendship links and community membership on LiveJournal, and the co-authorship network in the computer science bibliography hosted by the Digital Bibliography & Library Project (DBLP), showed that neither the threshold model nor the cascade model is sufficiently realistic (for more details, see in Section 2). Hence in this paper, we re-investigate the problem of trend adoption in social networks, and propose a new model that better predicts the observed behavior in real-life social networks.

Our first contribution in this paper is thus a new hybrid model for studying how individuals are influenced by their acquaintances in social networks. The hybrid network model considers the social network population as two groups: one group in which individuals follow cascade process to decide whether or not to adopt; the other group in which users obey threshold strategy to make the decision. The basic idea behind the hybrid network is that, with respect to trend adoption, we can distinguish between two types of individuals: 1) individuals who make “informed decisions” and 2) individuals who “imitate”. We refer to individuals who make informed decisions as *informed adopters*. Informed adopters have (insider) knowledge about the trend to adopt and decide whether or not to adopt based on information/discussion with other insiders who have already adopted. As a result, among the insider’s group, the adoption process can be modeled by the cascade process. We refer to individuals who imitate as *followers*. Followers do not have enough knowledge about the trend so as to make informed decisions, and as a consequence, decide whether or not to adopt depending on how many other people already have adopted the item. Hence for the latter group, the adoption process can be captured by the threshold model. In addition, we consider an initial set of individuals who adopt the trend without being influenced by others. We refer to these individuals as *trendsetters (early adopters)*. The trendsetters are important as they serve as “seeds” that enable the spread of the trend.

Note that followers have no special knowledge about the trend and therefore need to rely on their observation of what informed adopters do in order to decide on whether or not to adopt the trend. As such, the goal of a follower is to adopt the trend only if a large fraction of informed adopters does so, since this can be taken as an indication that the trend is indeed worth adopting. However, followers do not have a “global view” but only a “local view”, i.e. they can only observe what the informed adopters who are neighbors in their social network do, but they can not observe the adoption behavior of all informed adopters. As such, the question we want to explore is whether it is possible to choose a threshold value that guarantees that followers make the correct decision and adopt the trend only if a majority of the informed adopters does so, even though they have only local information.

Our second contribution is to characterize a *optimal threshold strategy* for the followers. More precisely we show that is optimal for followers to adopt a trend as soon as two friends in their social network adopted the trend. Interestingly, this behavior was observed in the Backstorm el. [4] on friendship links and community membership on LiveJournal, and the co-authorship network in DBLP.

To the best of our knowledge, it is the first time that a mathematical model is proposed to address and reassemble the real-life influence propagation pattern.

The rest of the paper is organized as follows. In Section 2, we discuss related work of empirical studies on real social networks. Section 3 presents our mathematical model, problem statement and numerical results. In Section 4, we propose a tractable model. In Section 5, we perform the analysis to address the choice of threshold value. Finally we conclude and discuss the future work in Section 6.

II. Related Work

In this section, we provide the related work of empirical studies concerning threshold strategy on real social networks. The subject of threshold is not new. In fact, it is well used in various fields of science [3]: as a method of segmentation in image processing, as a model in toxicology, as a phenomenon in particle physics, etc. However, the mathematical study of the threshold strategy and the optimum threshold value on social networks is the first time in the paper.

Backstorm el. [4] studied two large sources of data: friendship links and community membership on LiveJournal, and the co-authorship network in DBLP. They plotted the individual's probability p of joining a LiveJournal/DBLP community as a function of the number of friends, k, already in the community (Fig1, Fig2). Both plots exhibit qualitatively similar shapes in which p continues increasing, but more and more slowly as k increases. We take a closer look at Fig.1, since the error bars are smaller here. The probability of joining LiveJournal has a close fit to a logistic function for $k = 0, 1, 2$, with a rapid growth at $k = 2$. After this point, the probability of joining the community has a close fit to a logarithmic function. For the largest jump in p at $k = 2$, Backstorm el. claimed that the marginal benefit of having a second friend in a community is particularly strong. Such a curve suggests the individual's strategy of a threshold pattern: individuals have a low tendency to take an action if not many acquaintances did so; then the tendency increases very fast if a certain number of acquaintances have done so; beyond this point the tendency still increases, but does so slower and slower. The *action* herein is an abstract notation, which could be joining a community (Fig.1), or attending another conference (Fig.2). Though the empirical study of Backstorm el. reveals the individuals' threshold strategy, their research focus was on quite a different subject: the evolution of large social networks.

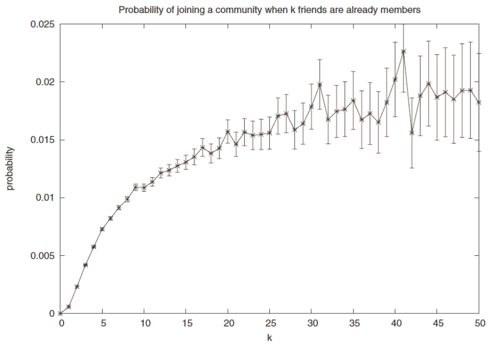

Fig. 1. The probability p of joining a LiveJournal community as a function of the number of friends k already in the community. Error bars represent two standard errors [4].

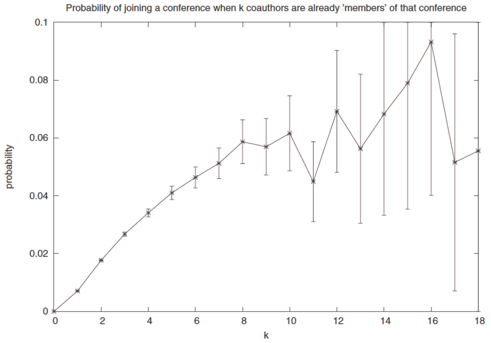

Fig. 2. The probability p of joining a DBLP community as a function of the number of friends k already in the community. Error bars represent two standard errors [4].

One variation of the above threshold strategy is in the context of voluntary vaccination [5]: once a sufficient proportion of the population is already immune, either naturally or by vaccination, people tend to skip vaccination under the voluntary policy due to morbidity risks of the vaccination itself. That is individuals tend to avoid taking an action once a threshold is reached. Bauch el. [5] performed a game theoretical analysis based on Nash equilibrium to explain the threshold strategy with respect to vaccination. However, they need to assume that all individuals are provided with the same information and use this information in the same way to assess risks.

Here this paper lifts the assumption of synchronized global information, and considers the situations where individuals have social contacts with their friends, thus have merely local observations. And individuals make decisions based on the limited information. To the best of our knowledge, our work provides a unique approach to choose a threshold value so that individuals can make correct choices based on local observations.

III. Model and Problem Statement

In this section, we introduce our hybrid model for trend adoption in social networks, state the question of interests, and provide a numerical case study to illustrate our result.

Recall that our scenario of trend adoption in the social network: individuals can observe whether their friends have adopted the trend or not, and make the decision based on these observations. We refer to an individual who adopted the trend as *active*, and to an individual who has not adopted the trend as *non-active*, or inactive. Note that individuals only go from non-active to active status, but not from non-active to active.

A. Social Network Graph

Our model consider two main groups in the social network population. One is the "informed adopter" group consisting of individuals who have knowledge about the trend and make informed decision based on information/discussion with other insiders who have already adopted. For this group, we model their user behaviour as cascade process. The other is the "follower" group consisting of individuals who don't have enough knowledge about the trend to make informed decisions, and as a consequence, decide whether or not to adopt by "imitating" their acquaintances who already have adopted the trend. For the second group, we presume they obey threshold strategy.

We think of a *social network* as a collection of individuals who have connections with one another in social bonds such as friendship, kinship, or relationships of beliefs and knowledge [7]. We use a graph G to represent the entire social network. As part of the social network graph G, we define a undirected subgraph G_1 of G representing the "informed adopters" and directed subgraph G_2 representing the "followers".

In the rest of the paper, we will use the terms individuals and nodes/vertices interchangeably.

More precisely, we use the following Erdös-Rényi [11] random graph model. We denote with $V(G_1)$ the vertex set of G_1 given by

$$V(G_1) = \{1, 2, 3, \cdots, n_1\},$$

where n_1 denotes the total number of informed adopters in the social graph G. Each node in $V(G_1)$ represents an informed adopter in the social network.

Between any pair of nodes $i, j \in V(G_1)$, with probability $p_1 \in (0, 1]$ there exists an undirected edge between two informed adopters i and j in G_1, independent of everything else. Let $E(G_1)$ be the resulting edge set of G_1 given by

$$E(G_1) = \{e_{i,j} | i, j \in V(G_1)\}.$$

Each edge in $E(G_1)$ stands for the social connection within G_1.

The average node degree λ_1 in G_1 is then given by

$$\lambda_1 = p_1 n_1$$

and we have the relation

$$p_1 = \frac{\lambda_1}{n_1}. \qquad (1)$$

Note that λ_1 represents the average number of informed adopters a node in G_1 communicates with.

Let $V(G_2)$ the vertex set of G_2 given by

$$V(G_2) = \{1, 2, 3, \cdots, n_2\},$$

where n_2 denotes the total number of followers in the social graph G. Each node in $V(G_2)$ represents a follower in the social network.

We assume that with probability $p_2 \in (0.1]$ there exists a directed edge $e(i, j)$ between a follower i in G_2 and an informed adopter j in G_1, independently of everything else.

Let $E(G_2)$ be the edge set of nodes in G_2. Each edge in $E(G_2)$ stands for the social connection between a follower in G_2 and an informed adopter in G_1, i.e. we have

$$E(G_2) = \{e_{i,j} | i \in V(G_1), j \in V(G_2)\}.$$

The average node degree λ_2 in G_2, i.e. the average number of informed adopters that a follower in G_2 knows, is then given by

$$\lambda_2 = p_2 n_1$$

and we have the relation

$$p_2 = \frac{\lambda_2}{n_1}.$$

Using the above notation, the graphs G_1 and G_2 are characterized by the parameters (n_1, λ_1) and (n_2, λ_2), respectively. Accordingly, we use the notation $G_1(n_1, \lambda_1)$ and $G_2(n_2, \lambda_2)$ to characterize the two graphs.

Aggregated, let $e(G)$ be the random variable denoting the number of edges in G.

$$e(G) = |E(G)|$$

And let $N(v)$ be the set of immediate *neighbours* of v, which stands for the set of friends, relatives or acquaintances of v on the social network.

$$N(v) = \{w | w \in V(G), \text{ and } e_{v,w} \in E(G)\}$$

B. Asymptotic Behavior

In the following we are interested in the asymptotic behavior as the social graph becomes large. As such, we consider an infinite sequence of social graphs denoted by $n = 1, 2, ...$, and let $G_1(n_1, \lambda_1)$ and $G_2(n_2, \lambda_2)$ be functions of n; i.e. the parameters (n_1, λ_1) and (n_2, λ_2) of G_1 and G_2, respectively, depend on n. We assume that n_1 and n_2 become large (approach infinity) as n increases.

Assumption 1. *We have*

$$\lim_{n \to \infty} n_1(n) = \infty$$

and

$$\lim_{n \to \infty} n_2(n) = \infty.$$

To keep the notation light, in the following we typically do not explicitly state the dependency of parameters such as (n_1, λ_1) and (n_2, λ_2) on n.

Note that the above implies that the edge probabilities p_1 and p_2 are also functions of n. For our analysis, we make the following assumption on $p_2(n)$.

Assumption 2. *We have*

$$\lim_{nto\infty} p_2(n) = \frac{\lambda_2(n)}{n_1(n)} = 0.$$

The above assumption states that the followers can only observe a vanishingly small fraction of the informed adopters.

C. Trendsetters

In addition to the informed adopters and followers, we consider a third group of individuals, namely the trendsetters or early adopters. Trendsetters adopt the trend without being influenced by others in the social network. Trendsetters serve as "seeds" that enable the spread of the trend.

In our model, the trendsetters are given by a subset A_1 of the informed adopters G_1, i.e. A_1 is the the initial active set within the insiders group, G_1. This size of the set A_1 can depend on n and we make the following technical assumption on the size of the set $A_1(n)$, i.e. the total number of trendsetters as a function of n.

Assumption 3. *There exist positive constants ϵ and k, and $c \in [0, \frac{1}{2})$ such that*

$$\lim_{n\to\infty} |A_1(n)| > \epsilon$$

and

$$\lim_{n\to\infty} |A_1(n)| \leq k * (n_1^c). \tag{2}$$

We restrict the size of $A_1(n)$ because we want to focus on the case where the numbers of trendsetters does not dominate the total number of informed adopters.

D. Trend Adoption

We consider the following policies for informed adopters and followers to adopt a trend.

Informed Adopters – Influence Cascade: Informed adopters in G_1 are influenced by the decisions made by other informed adopters who are within one hop friendship on $G_1(n_1, \lambda_1)$. We model this process using the *independent information cascade* model that was introduced to study the problem of the maximum influenced set in social networks [6], [8], [9].

More precisely, we divide the set $E(G_1)$ into two disjoint subsets of *open* and *blocked* edges. An edge $e(i,j) \in E(G_1)$ is open with probability $\rho_1 \in [0,1]$, independent of everything else; otherwise it is blocked.

The trend adoption then only propagates over open edges, i.e. if node i has adopted the trend and node $e(i,j)$ is an open edge, then node j will also adopt the trend [6], [8], [9].

Followers – Threshold Strategy: Here we assume that followers use a *threshold policy* in order to adopt the trend, i.e. a follower will adopt the trend only if the number of informed adopters that adopted the trend exceeds a given threshold value. More precisely, let t_a be a threshold value used by the followers in $G_2(n_2, \lambda_2)$ in order to decide whether or not to adopt the trend. A follower then adopts the trend only if he observes t_a number of adoptions among his one-hop friends.

E. Adoption Process

The trend adoption process then proceeds as a discrete time. At time $t = 0$, only the trendsetters A_1 in G_1 have adopted the trend, i.e. are active in G_1. The trend then spreads from A_1 as follows. When a node $i \in V(G_1)$ first becomes active in step t, it is given a single chance to activate each currently non-active neighbour $j \in N(i)$, and it succeeds with probability ρ_1. If i succeeds, then j will become active in step $t+1$. The cascade terminates when no node has become newly active at a given time step. Followers $V(G_2)$ in G_2 observe whether their one-hop friends in G_1 and adopt the trend as soon as t_a of their friends have adopted the trend, i.e .have become active.

F. Problem Statement

In the above scenario of the individual adoptions on social networks, we are interested in the question of an optimal threshold decision policy, which allows followers to make the correct decision in the sense that they only adopts a trend if the majority of informed adopters in G_1 will adopt the trend. Thus we are interested in which threshold value t_a to choose such that followers will almost always make the correct decision to conform with the majority of individuals in G_1. Note that the threshold decision policy merely makes use of individuals' local observations.

It is well known [11] that (asymptotically) a majority of informed adopters in G_1 will adopt the trend only if

$$\lim_{n\to\infty} \rho_1 \lambda_1 > 1$$

and only a small fraction of informed adopters in G_1 will adopt the trend if

$$\lim_{n\to\infty} \rho_1 \lambda_1 < 1.$$

Therefore, the question we want to study is whether there exists a threshold value t_a that ensures that (asymptotically) no follower will adopt the trend in the case where $\lim_{n\to\infty} \rho_1 \lambda_1 < 1$, and each follower will adopt the trend with positive probability if $\lim_{n\to\infty} \rho_1 \lambda_1 > 1$.

G. Main Result

Our main result is to prove that choosing the threshold value $t_a = 2$ is optimal for followers in the sense that the threshold value $t_a = 2$ ensures that a follower never makes a mistake by adopting a trend that is adopted only by a small fraction of the informed adopters, while maximizing the probability that a follower adopts the trend when a majority of the informed adopters does so.

More precisely, we obtain the following result. Consider a follower i in G_2 and let the random variable K_i denote the number of active neighbours of i after the cascade process

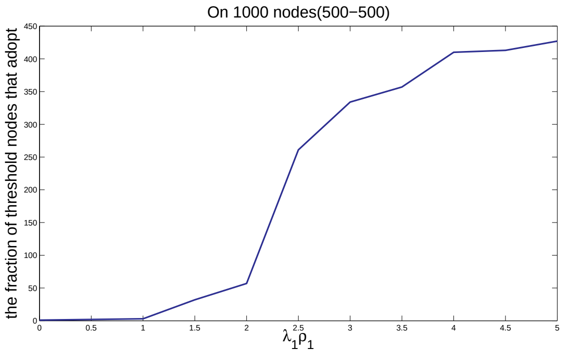

Fig. 3. Double jump at $\lambda_1\rho_1 = 1$.

has finished, i.e. the numbers of active informed adopters in G_1 to whom follower i is connected in the social graph G. Conditional on $\lambda_1\rho_1$, let $F(k)$ be the probability there exists at least one node in G_2 who has at least k number of active neighbours in G_1, i.e. $F(k)$ is given by

$$F(k) = Pr(\exists i \in V(G_2),\ \text{such that } K_i \geq k|\lambda_1\rho_1 < 1). \quad (3)$$

In our main result in Section 5 we show that

$$\lim_{n\to\infty} F(1) > 0$$

and

$$\lim_{n\to\infty} F(k) = 0, \quad \text{for } k = 2, 3, \cdots .$$

Note that this result implies that under a any threshold value $t_a \geq 2$, no follower will make the mistake to adopt a trend if only a small fraction of the informed adopters does so. This is not true for the threshold value $t_a = 1$, i.e. under the threshold value $t_a = 1$ some followers will make a the mistake and adopt the trend even though only a small fraction of the informed adopters does so. Hence, using a threshold $t_a \geq 2$ is “safe” as ensures that followers never make a mistake. Moreover, among all value $t_a \geq 2$ the value $t_2 = 2$ is optimal in the sense that it maximizes the probability of adopting the trend when a majority of the informed adopters does so. In particular, one can show that under the threshold $t_a = 2$ followers will adopt the trend with a strictly positive probability when a majority of the informed adopters adopt the trend.

H. Numerical Results

To illustrate our result, we provide a numerical case study. First, we plot the fraction of nodes in G_2 that adopts the trend as the product of $\lambda_1\rho_1$ increases (Fig. 3). This leads to an important observation, known as *the double jump* [12].

Furthermore, in order to examine closely the relation between individual adoption behaviour and their neighbourhood adoption status, we plot Fig. 4 , in which we find that the fraction of nodes who adopt the trend has a sudden rise with two active neighbours presenting. Interestingly, this observation was experimentally observed in real social networks [4]. This is one important feature of our model. As before, no propagation models can reproduce this “second-adoption” phenomenon (Fig.1, 2) that appears in real life.

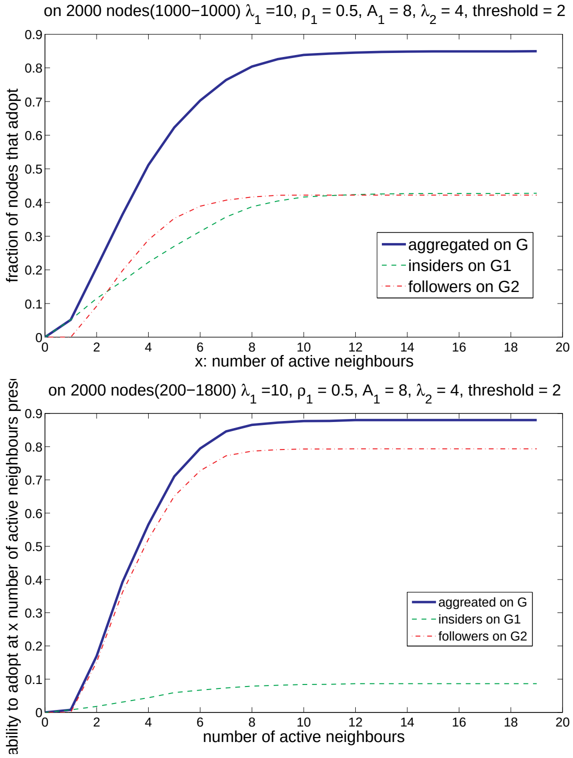

Fig. 4. Second-adoption phenomenon.

IV. Equivalent Model

In this section, we propose the equivalent model of G in Section 3, $G'(n_1, n_2, \lambda_1\rho_1, A_1, \lambda_2, t_a)$. In particular, we shall prove that the size of the influenced set produced by ρ_1-cascade process over $G_1(n_1, \lambda_1, A_1)$ has the same distribution as that of the influenced set on $G'_1(n_1, \lambda_1\rho_1, A_1)$.

Before proceeding, we need to define the *influenced set* I_1 on G_1. Let the influenced set be the set of vertices connected to A_1 via open edges in G_1, i.e.,

$$I_1 = \{v|v \in V(G_1), v \text{ connects to } A_1 \text{ via } \textit{open} \text{ edges}\}$$

Hence by the end of the influence cascade process, all the nodes in I_1 become active.

The equivalent model G'_1 is a Erdös-Rényi graph. G'_1 has the same vertex set as G_1, but with a different edge preserving probability

$$p'_1 = p_1\rho_1,$$

where p_1 is defined in (3.1), the edge preserving probability of G_1, the success probability of the influence cascade process. G'_1 have the same initial active set as G_1, A_1, and subjecting to the same size constraint.

Now we define the *influenced set* I'_1 on G'_1. Let the influenced set on G'_1 be the set of vertices connected to A_1 via edges in G'_1, i.e.,

$$I'_1 = \{v | v \in V(G'_1), v \text{ connects to } A_1 \text{ via edges } e \in E(G'_1)\}$$

The following theorem guarantees that on both models the distribution in the size of the influenced set is the same. To prove that unfolding the influence cascade by determining the open/block status of edges gradually at each time step is the same as unfolding the influence cascade by determining the status of all the edges at the beginning during the construction of G'_1, the idea is to argue that the distribution for a random node joining the influenced set is the same over time steps in both models.

Theorem 1. *Given an initial active set A_1, the distribution of the influenced sets obtained by cascade process on G_1 starting from A_1, is the same as the distribution of sets reachable from A_1 via edges on G'_1.*

Proof: First we argue the iterations of the influence cascade process over G_1. Define $A_1^{(t)}$ to be the set of active nodes at the end of iteration t, for $t = 0, 1, 2, 3, \cdots$. If a random node v has not become active by the end of iteration t, then under the cascade process, the probability that v will become active in iteration $t+1$ is equal to the chance that one of its active neighbours in $A_1^{(t)}$ but not in $A_1^{(t-1)}$ succeeds in activating v. This probability is $P(v \in A_1^{(t+1)} | v \notin A_1^{(t)}) = 1 - \prod_{w \in A_1^{(t)} \setminus A_1^{(t-1)}} (1 - \frac{\rho_1 \lambda_1}{n_1})$.

Second we construct G'_1 gradually as follows. We start with the initial set A_1. For each node v with at least one edge stub, we determine whether it connects to A_1. If so, then v is reachable; if not, we keep the source of v's edge unknown subject to the condition that it comes from outside of A_1. Having now exposed a new set of reachable nodes $A_1^{(1)}$ in the first time step, we proceed to reveal further reachable nodes by performing the same process on edges from A_1, and in this way produce sets $A_1^{(2)}, A_1^{(3)}, A_1^{(4)}, \cdots$. If node v has not been determined to be reachable by the end of time step t, then the probability that it is determined to be reachable in time $t+1$ is equal to the chance that its edge comes from $A_1^{(t)}$ but not in $A_1^{(t-1)}$; this probability is $A_1^{(t+1)} | v \notin A_1^{(t)}) = 1 - \prod_{w \in A_1^{(t)} \setminus A_1^{(t-1)}} (1 - \frac{\rho_1 \lambda_1}{n_1})$.

Thus, by induction over the iterations, we see that the cascade process produces the same distribution over influenced sets as the construction of G'_1. ■

Since we have established the equivalence between the $G_1(n_1, \lambda_1, \rho_1, A_1)$ in Section 3 and $G'_1(n_1, \lambda_1 \rho_1, A_1)$ with respect to the distribution in the size of the influenced set, and G'_1 is essentially a Erdös-Rényi graph.

V. Analysis

In this section, we derive in detail the answer to the question of the optimal trend-adoption policy on $G'(n_1, n_2, \lambda_1 \rho_1, A_1, \lambda_2, t\_a)$. We shall specifically consider the adoption policy in subcritical phase, and obtain the optimal policy.

A. Decision Policy in Subcritical Phase

When $p'_1 = p_1 \rho_1 \in [0, \frac{1}{n_1})$, G'_1 is in subcritical phase. We first derive the probability, $u(k)$, for a random vertex in $G_2(n_2, \lambda_2)$ possessing k active neighbours. Then by making use of this formula, we derive the conditional probability, $F(k)$, for existing any node in $G_2(n_2, \lambda_2)$ with at least k active neighbours. And we thus prove that in the subcritical phase the correct decision policies are threshold strategies with threshold value: $t_a = 2, 3, \cdots$.

Let P_a be the probability for a random node in G'_1 being active after the influence cascade terminates. Then P_a is given by:

$$P_a = \frac{|I_1|}{n_1} \tag{4}$$

In order to quantify P_a, we make use of the expectation of $|I'_1|$, which can be expressed in closed form using percolation theory on random graphs. Percolation property of random graphs has been extensively studied in the field of epidemics spread [15], [16], [17]. It studies whether a disease will dominate the whole population. If the disease does not percolate the whole network, it quantifies the expected value of the size of the infected population. The main approach used is generating functions [18]. By the same approach, we are now deriving $E(|I'_1|)$ on Erdös-Rényi random graph, $G_{n,p}$, with percolation probability ρ.

Theorem 2. *On Erdös-Rényi random graph $G(n, p = \frac{\lambda}{n})$, the expectation of the size of the influenced set with one vertex percolated initially is $\frac{1-\rho}{1-\rho-\lambda\rho}$.*

Proof: Let p_k be the probability for a random vertex having degree k. p_k is given by:

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k} \approx \frac{\lambda^k e^{-\lambda}}{k!}. \tag{5}$$

Let $G_0(x)$ be the generating function of $\{p_k\}$, which is given by:

$$G_0(x) = \sum_{k \geq 0} p_k x^k$$

Follow a random edge and reach a vertex v. Let $\{q_k\}$ be the degree distribution of so-reached vertex v. Let $G_1(x)$ be the generating function of $\{q_k\}$, which is given by:

$$G_1(x) = \sum_{k \geq 0} q_k x^k$$

A random edge has probability $\frac{1}{e(G)}$ to be chosen by the uniform choosing policy, where $e(G)$ denotes the total number of edges in G. Conditional on v having degree k, the chosen edge has probability $\frac{k}{e(G)}$ being incident to v, i.e.,

$$P(\text{ the chosen edge incident to } v | deg(v) = k) = \frac{k}{e(G)} \tag{6}$$

By Total law of probability, from (5, (6) we have:

$$P(\text{the chosen edge incident to } v) = \frac{\sum_k p_k k}{e(G)}. \tag{7}$$

Thus by Bayes rule, from (5), (7) we have q_k which is given by:

$$\begin{aligned} q_k &= P(deg(v) = k| \text{ the chosen edge incident to } v) \\ &= \frac{p_k k}{\sum_k p_k k} = \frac{p_k k}{z} \end{aligned} \tag{8}$$

where z is the average vertex degree. And in the case of Erdös-Rényi random graph $z = \lambda$ on our model.

Let $\tilde{p}_m$ be the probability for a random vertex having m open edges. Let $G_0(x;\rho)$ be the generating function of $\{\tilde{p}_m\}$, which is given by:

$$\begin{aligned} & P(m \text{ edges of } v \text{ are open } |deg(v) = k) \\ & = \binom{k}{m} \rho^m (1-\rho)^{k-m} \text{ binomial distribution} \\ & \Rightarrow \\ & \tilde{p}_m = P(m \text{ edges of } v \text{ are open}) \\ & = \sum_k p_k \binom{k}{m} \rho^m (1-\rho)^{k-m} \\ & \Rightarrow \\ & G_0(x;\rho) = \sum_{m \geq 0} \sum_{k \geq 0} p_k \binom{k}{m} \rho^m (1-\rho)^{k-m} x^m \\ & = \sum_{k \geq 0} p_k \sum_{m \geq 0}^{k} \binom{k}{m} (x\rho)^m (1-\rho)^{k-m} \\ & = \sum_{k \geq 0} p_k (1 - \rho + x\rho)^k \end{aligned}$$

Thus

$$G_0(x;\rho) = G_0(1 - \rho + x\rho) \tag{9}$$

Now follow a random edge and reach a vertex v. Let $\tilde{q}_m$ be the probability for so-reached v having m open edges. Let $G_1(x;\rho)$ be the generating function of $\{\tilde{q}_m\}$. Similar with the derivation of (9), $G_1(x;\rho)$ is given by:

$$G_1(x;\rho) = G_1(1 - \rho + x\rho) \tag{10}$$

Let $\mathbf{Z}$ be a random variable denoting the size of influenced set connected to a random *vertex*. Define $H_0(x;\rho)$ be generating function for the distribution of $\mathbf{Z}$. Let $\tilde{Z}$ be a random variable denoting the size of influenced set connected to a random *edge*. Define $H_1(x;\rho)$ be generating function for the distribution of $\tilde{Z}$. The influenced components generated by $H_0(x;\rho)$ consist of the initial infected node, plus any number of tree-like clusters, joined to it by single edges. Each tree-like cluster has the size distribution generated by $H_1(x;\rho)$. Thus we have:

$$\begin{aligned} H_0(x;\rho) = & \ xq_0 + xq_1 H_1(x;\rho) \\ & + xq_2[H_1(x;\rho)]^2 + xq_3[H_1(x;\rho)]^3 + \dots \end{aligned}$$

i.e.,

$$H_0(x;\rho) = xG_0(H_1(x;\rho);\rho)$$

and

$$H_1(x;\rho) = x\tilde{q}_0 + x\tilde{q}_1 H_1(x;\rho) + x\tilde{q}_2[H_1(x;\rho)]^2 + \dots$$

i.e.,

$$H_1(x;\rho) = xG_1(H_1(x;\rho);\rho) \tag{11}$$

The expected value of $\mathbf{Z}$ is given by differentiating $H_0(x;\rho)$, i.e.,

$$E[\mathbf{Z}] = H_0'(1;\rho) = 1 + G_0'(1;\rho)H_1'(1;\rho) \tag{12}$$

Differentiating equation (11), we have:

$$H_1'(1;\rho) = 1 + G_1'(1;\rho)H_1'(1;\rho) = \frac{1}{1 - G_1'(1;\rho)} \tag{13}$$

substitute (13) into (12), and make use of (8),(9), (10), we have:

$$E[\mathbf{Z}] = 1 + \frac{\rho G_0'(1)}{1 - \rho G_1'(1)} \tag{14}$$

We notice

$$G_0'(1) = \sum_{k \geq 0} k p_k x^{k-1}|_{x=1} = \sum_{k \geq 0} k p_k = \lambda \tag{15}$$

With (8), we also have

$$\begin{aligned} G_1'(1) &= \sum_{k \geq 0} k q_k x^{k-1}|_{x=1} \\ &= \sum_{k \geq 0} \frac{k^2 p_k}{z} = \frac{E(K^2)}{z} \\ &= \frac{Var(K) + E^2(K)}{z} = \frac{\lambda + \lambda^2}{\lambda} \\ &= 1 + \lambda \end{aligned} \tag{16}$$

where K in (16) is a r.v. denoting the vertex degree of a random node in Erdös-Rényi random graph $G(n, p = \frac{\lambda}{n})$. Thus, K has approximately Poisson distribution with $E(K) = \lambda$ and $Var(K) = \lambda$. Substituting (15) and (16) into (14), we have

$$E[\mathbf{Z}] = \frac{1 - \rho}{1 - \rho(1 + \lambda)}$$

■

By Theorem 2, we have the expectation of the influenced set given by:

$$E(|I_1'|) = \frac{1 - \rho_1}{1 - \rho_1(\lambda_1 + 1)} \text{ for } |A_1| = 1$$

Further for $|A_1| \leq k * n_1^c$ for some k, we have

$$E(|I_1'|) \leq \frac{1 - \rho_1}{1 - \rho_1(\lambda_1 + 1)} k * n_1^c = \leq k' * n_1^c, \text{ for some } k'.$$

If the influenced sets associated with each node in A_1 are independent, the sum of the influenced sets has the expectation of $\frac{1-\rho_1}{1-\rho_1(\lambda_1+1)} k * n_1^c$. The inequality comes from the possibility that those influenced sets overlap with one another.

By Markov inequality and Theorem 2, we have an upper bound of $|I_1'|$ given by:

$$Pr(|I_1'| \geq n^{2c}) \leq \frac{E(|I_1'|)}{n_1^{2c}} \leq \frac{kn_1^c}{n_1^{2c}} = \frac{k}{n_1^c} \to 0 \text{ for } n_1 \text{ large}$$

Since we have restricted the size of A_1 by assuming that $0 \leq c < \frac{1}{2}$ in (2), then $\lim_{n_1 \to \infty} P(|I'| \geq n_1^{c'}) = 0$, where $c' \in [0, 1)$. By Theorem 1, we have the same bound of $|I_1|$:

$$\lim_{n \to \infty} P(|I_1| \geq n_1^{c'}) = 0, \text{ where } c' \in [0, 1) \qquad (17)$$

Alternatively, (17) can be written as:

$$\lim_{n_1 \to \infty} P(|I_1| < n_1^{c'}) = 1 \qquad (18)$$

Now we bound P_a by substituting (18) into (4):

$$P_a = \frac{|I_1|}{n_1} < \frac{n_1^{c'}}{n_1} = 0 \text{ with probability 1 for } n_1 \text{ large} \qquad (19)$$

Equation (19) has an intuitive interpretation. In the subcritical phase of $G_1'(n_1, \lambda_1\rho_1, A_1)$, we randomly pick a node when the influence cascade terminates. Such a node is non-active with probability 1 for large networks. The following theorem provides the distribution for a random node in $G_2(n_2, \lambda_2)$ having n_a number of active neighbours.

Theorem 3. *On our model $G'(n_1, n_2, \lambda 1\rho_1, A_1, \lambda_2, t_a)$, we have $P(K_v = n_a) = \frac{(\lambda_2 P_a)^{n_a}}{n_a! e^{\lambda_2 P_a}}$, where P_a is given by (4).*

Proof: First, we derive the degree distribution of a random node in $G_2(n_2, \lambda_2)$. Let p_k be the probability for a random node $v \in G_2$ having k neighbours. p_k satisfies binomial distribution and is given by:

$$p_k = \binom{n_1 - 1}{k} p_2^k (1 - p_2)^{n_1 - 1 - k}$$

where $p_2 = \frac{\lambda_2}{n_1}$.

We apply Poisson approximation as n_1 goes to infinity:

$$p_k = \binom{n_1 - 1}{k} p_2^k (1 - p_2)^{n_1 - 1 - k} \approx \frac{\lambda_2^k e^{-\lambda_2}}{k!} \qquad (20)$$

Next we derive the distribution of the number of active neighbours a random node in G_2 possess. On the condition that a random node v has degree k, the probability for v having n_a active neighbours also satisfies binomial distribution:

$$P(K_v = n_a | deg(v) = k) = \binom{k}{n_a} P_a^{n_a} (1 - P_a)^{k - n_a} \qquad (21)$$

where P_a is the probability for a random node being active, as defined in (4).

By Total law of probability, we remove the condition in (21), and have the probability for v having n_a active neighbours given by:

$$P(K_v = n_a) = \sum_k p_k \binom{k}{n_a} P_a^{n_a} (1 - P_a)^{k - n_a} \qquad (22)$$

By substituting p_k in (20) into (22) and applying Taylor expansion [20], we have

$$\begin{aligned} P(K_v = n_a) = & \\ &= \sum_k \frac{\lambda_2^k e^{-\lambda_2}}{k!} \frac{k!}{(k - n_a)! n_a!} P_a^{n_a} (1 - P_a)^{k - n_a} \\ &= \frac{e^{-\lambda_2} P_a^{n_a}}{n_a! (1 - P_a)^{n_a}} \sum_k \frac{\lambda_2^k (1 - P_a)^k}{(k - n_a)!} \\ &= \frac{e^{-\lambda_2} P_a^{n_a}}{n_a! (1 - P_a)^{n_a}} [\lambda_2 (1 - P_a)]^{n_a} e^{\lambda_2 (1 - P_a)} \\ &= \frac{(\lambda_2 P_a)^{n_a}}{n_a! e^{\lambda_2 P_a}} \end{aligned}$$

■

Next, we use Theorem 3 to determine threshold decision policy. In particular, we define $F(k)$ to be a conditional probability. It denotes the probability that there exists one G_2 node having at least k number of active neighbours, conditional in subcritical phase. The next theorem quantifies $F(k)$, which in turn suggests that if choosing $t_a \geq 2$, with probability 1, no nodes in G_2 make a wrong decision.

Theorem 4. *We have*

$$\lim_{n \to \infty} F(1) > 0$$

and

$$\lim_{n \to \infty} F(k) = 0, \quad \textit{for } k = 2, 3, \cdots.$$

Proof: To see whether there exists a "follower" node in G_2 having at least one active insider-neighbour in subcritical phase, we compute $\lim_{n \to \infty} F(1)$:

$$\begin{aligned} \lim_{n \to \infty} F(1) &= \lim_{n \to \infty} 1 - [1 - Pr(K_v \geq 1)]^{n_2} \\ &= 1 - \lim_{n \to \infty} [Pr(K_v = 0)]^{n_2} \\ &= 1 - \lim_{n \to \infty} [\frac{(\lambda_2 P_a)^0}{0! e^{\lambda_2 P_a}}]^{n_2} \\ &= 1 - \lim_{n \to \infty} [\frac{(\lambda_2 \frac{|I_1|}{n_1})^0}{0! e^{\lambda_2 \frac{|I_1|}{n_1}}}]^{n_2} \\ &= 1 - \frac{1}{e^{\lambda_2 |I_1| \frac{n_2}{n_1}}} > 0 \end{aligned}$$

To see whether there exists a node in G_2 having at least two active neighbours in subcritical phase, we compute $\lim_{n \to \infty} F(2)$:

$$\begin{aligned} \lim_{n \to \infty} F(2) &= \lim_{n \to \infty} 1 - [1 - Pr(K_v \geq 2)]^{n_2} \\ &= 1 - \lim_{n \to \infty} [Pr(K_v = 0) + Pr(K_v = 1)]^{n_2} \\ &= 1 - \lim_{n \to \infty} [\frac{(\lambda_2 P_a)^0}{0! e^{\lambda_2 P_a}} + \frac{(\lambda_2 P_a)^1}{1! e^{\lambda_2 P_a}}]^{n_2} \\ &= 1 - \lim_{n \to \infty} [\frac{(\lambda_2 \frac{|I_1|}{n_1})^0}{0! e^{\lambda_2 \frac{|I_1|}{n_1}}} + \frac{(\lambda_2 \frac{|I_1|}{n_1})^1}{1! e^{\lambda_2 \frac{|I_1|}{n_1}}}]^{n_2} \\ &= 1 - \lim_{n \to \infty} \frac{(1 + \frac{\lambda_2 |I_1|}{n_1})^n}{e^{\lambda_2 |I_1|}} \end{aligned} \qquad (23)$$

To further reduce (23), let $m = \frac{n_1}{\lambda_2|I_1|}$. Note that $\lim_{n\to\infty} m = \infty$. Thus we reduce the nominator in (23) as follows:

$$\begin{aligned} &\lim_{n\to\infty}(1+\frac{\lambda_2|I_1|}{n_1})^{n}_{1} \\ &= \lim_{n\to\infty}\{(1+\frac{\lambda_2|I_1|}{n_1})^{\frac{n_1}{\lambda_2|I_1|}}\}^{\lambda_2|I_1|} \\ &= \lim_{m\to\infty}[(1+\frac{1}{m})^{m}]^{\lambda_2|I_1|} \\ &= e^{\lambda_2|I_1|} \end{aligned} \tag{24}$$

In the derivation of (24), we make use of the common limit [21]:

$$\lim_{m\to\infty}(1+\frac{1}{m})^{m} = e$$

Substitute (24) into (23). We have:

$$\begin{aligned} \lim_{n\to\infty} F(2) &= 1 - \lim_{n\to\infty}\frac{(1+\frac{\lambda_2|I_1|}{n_1})^{n_2}}{e^{\lambda_2|I_1|}} \\ &= 1 - \frac{e^{\lambda_2|I_1|}}{e^{\lambda_2|I_1|}} \\ &= 0. \end{aligned} \tag{25}$$

(25) says that there exists no node having 2 or more active neighbours in subcritical phase. It generalizes that there exists no node having 3 or more active neighbours in subcritical phase, and so forth. Thus we have $\lim_{n\to\infty} F(k) = 0$, for $k = 2, 3, \cdots$. ■

Theorem 4 suggests the decision policy for followers in subcritical phase. The first part of the theorem says that the threshold value should not be $t_a = 1$. Since $\lim_{n\to\infty} F(1) > 0$ suggests that the probability is positive for existing at least one node having 1 active neighbour by the end of influence cascade process. If we choose $t_a = 1$ as the threshold value, then by the strategy the nodes with one active neighbour shall adopt thus make the wrong decision in subcritical phase. The second part says that the threshold value could be $t_a = 2, 3, \cdots$. Since $\lim_{n\to\infty} F(k) = 0$ for $k = 2, 3, \cdots$ says the probability tends to zero for existing any nodes having k or more active neighbours. In words, with probability one there is no node with k or more active neighbours for $k \geq 2$. Hence these choices of k can guarantee all the "follower" nodes in G_2 not to adopt thus make the correct decision in the subcritical phase. Moreover, one can show that under the threshold $t_a = 2$ followers will adopt the trend with a strictly positive probability when a majority of the informed adopters adopt the trend (i.e. when $\lim_{n\to\infty} \lambda_1\rho_1 > 1$); and the larger the fraction of informed adopters who adopted the trend, the higher the probability that a follower will do so. In this sense, the threshold value $t_a = 2$ the optimal value.

VI. Conclusions

In this paper, we study how an individual in a social network should decide whether or not to adopt a trend, based on how many people in their neighborhood in the social network adopted the trend. In particular we investigated the question in what adoption policy leads to an optimal trend-adoption in the sense that an individual only adopts a trend if the majority of their social network will do so. we obtain the result that the optimal policy for an individual is to adopt a trend if two people in their neighborhood did.

Our analysis is mainly done on the Erd$\ddot{o}$s-R$\acute{e}$nyi random graph model. Though our work is based on an oversimplified model, it indeed shed some light on how individuals make choices in real life. this result/behavior was experimentally observed in real social networks. The work of Backstrom el. [4] noticed the special effect of the second friend and claimed that the marginal benefit of having a second friend in the community is particularly strong. Here in our paper, we prove the optimality of the threshold value of 2, which coincides with the discovery of the special second adoption found in [4].We hope that this work will help towards building applications that is able to automatically push relevant content to users in online social networks.

References

[1] http://en.wikipedia.org/wiki/Conformity. access on March. 15th, 2010.

[2] Aronson, E., Wilson, T.D., Akert, A.M. Social Psychology(6th edition). Upper Saddle River, NJ: Pearson Prentice Hall. 2007.

[3] http://en.wikipedia.org/wiki/Threshold. access on March. 18th, 2010.

[4] L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan. Group Formation in Large Social Networks: Membership, Growth, and Evolution. Proc. 12th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2006.

[5] C.T. Bauch and D.J.D. Earn. Vaccination and the theory of games. Proceedings of the National Academy of Sciences 101: 13391-13394. 2004.

[6] D. Kempe, J. Kleinberg, E. Tardos. Maximizing the Spread of Influence through a Social Network. Proc. 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2003.

[7] D. Easley, J. Kleinberg. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. To be published by Cambridge University Press, 2010.

[8] D. Kempe, J. Kleinberg, E. Tardos. Influential Nodes in a Diffusion Model for Social Networks. Proc. 32nd International Colloquium on Automata, Languages and Programming (ICALP), 2005.

[9] J. Kleinberg. Cascading behaviour in Networks: Algorithmic and Economic Issues. In Algorithmic Game Theory (N. Nisan, T. Roughgarden, E. Tardos, V. Vazirani, eds.), Cambridge University Press, 2007.

[10] Hacking, Ian. Logic of statistical inference. Cambridge, University Press, 1965.

[11] B. Bollobs. Random graphs(2nd edition). Cambridge University Press, 2001.

[12] P. Erdos, A. Renyi. On the Evolution of Random Graphs. Publ. Math. Inst. Hungar. Acad. Sci. 5(1960), 17-61.

[13] R. Durrett. Random graph dynamics. Cambridge University Press, 2007.

[14] J. Spencer. The strange logic of random graphs. Springer, c2001.

[15] M. E. J. Newman. The spread of epidemic disease on networks. Phys. Rev. E 66, 016128. 2002.

[16] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. Phys. Rev. E 64, 026118. 2001.

[17] M. E. J. Newman. Random graphs with clustering. Phys. Rev. Lett. 103, 058701. 2009.

[18] Herbert S. Wilf. Generating functionology. Wellesley, Mass. : A K Peters, 3rd ed. 2006.

[19] NIST/SEMATECH, '6.3.3.1. Counts Control Charts', e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/pmc/section3/, March 2010.

[20] Greenberg, Michael, Advanced Engineering Mathematics (2nd ed.), Prentice Hall, ISBN 0-13-321431-1, 1998.

[21] Catherine Roberts, Ray McLenaghan. Continuous Mathematics in Standard Mathematical Tables and Formulae. ed. Daniel Zwillinger. Boca Raton: CRC Press (1996): 333, 5.1 Differential Calculus