

Singapore Management University Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems

School of Information Systems

6-2012

Modeling diffusion in social networks using network properties

Duc Minh LUU

Ee Peng LIM

Singapore Management University, eplim@smu.edu.sg


Tuan Anh HOANG

Singapore Management University, tahoang.2011@smu.edu.sg

Chong Tat Freddy CHUA

Singapore Management University, freddy.chua.2009@smu.edu.sg

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the [Databases and Information Systems Commons](#), and the [Numerical Analysis and Scientific Computing Commons](#)

Citation

LUU, Duc Minh; LIM, Ee Peng; HOANG, Tuan Anh; and CHUA, Chong Tat Freddy. Modeling diffusion in social networks using network properties. (2012). *Proceedings of the Sixth International Conference on Weblogs and Social Media*. Research Collection School Of Information Systems.

Available at: https://ink.library.smu.edu.sg/sis_research/1545

This Conference Proceeding Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

Modeling Diffusion in Social Networks using Network Properties

Minh-Duc Luu, Ee-Peng Lim, Tuan-Anh Hoang, Freddy Chong Tat Chua

Singapore Management University
80 Stamford Road, Singapore 178902

Abstract

Diffusion of items occurs in social networks due to spreading of items through word of mouth and exogenous factors. These items may be news, products, videos, advertisements or contagious viruses. When a user purchases or consumes one of such items, we say that she adopts the item and she becomes an item adopter. Previous research has studied diffusion process at both the *macro* and *micro* levels. The former models the number of item adopters in the diffusion process while the latter determines which individuals adopt item. Both macro and micro level models have their merits and limitations. In this paper, we establish a general probabilistic framework, which can be used to derive macro-level diffusion models, including the well known **Bass Model** (BM). Using this framework, we develop several other models considering the social network's degree distribution coupled with the assumption of *linear influence* by neighboring adopters in the diffusion process. Through some evaluation on synthetic data, this paper shows that degree distribution actually changes during the diffusion process. We therefore introduce a **multi-stage diffusion model** to cope with variable degree distribution. By conducting experiments on both synthetic and real datasets, we show that our proposed diffusion models can recover the diffusion parameters from the observed diffusion data, which allows us to model diffusion with high accuracy.

1 Introduction

1.1 Motivation

Diffusion research has its origin from the marketing science, social science and health science communities which focus on studying diffusion of innovation (including products), social behaviors, and epidemics respectively. These research resulted in a wide range of diffusion models that help to describe the rate of diffusion and number of adopters throughout the diffusion process. More recently, the availability of social networks and user transactional data have increased the importance of marketing through online social media. The need for more elaborate marketing techniques have motivated research on diffusion of information and product items in social networks.

The existing diffusion research can be broadly divided into **macro-** and **micro-**level studies. The macro-level studies model the aggregated number of adoptions in the user population throughout the diffusion process. The topics relevant to macro-level study include rate of adoption, number of adopters at any time point, and peak of adoptions. One may assume some or no knowledge about the relationships among users. At one extreme, each user is assumed to be related to every other users, i.e., a complete network. Based on this assumption, the well known **Bass Model** (Bass 1969) and other similar macro-level models have been well studied in the literature and applied to many different diffusion scenarios and applications, (Meade and Islam 2006), (Iribarren and Moro 2011). At the other extreme, users are connected by some social or communication network (usually a non-complete network) which allows items to be diffused to users only through *neighboring* adopters. Nevertheless, there are very few works on developing macro-level models which consider such kind of networks, (Shaikh, Rangaswamy, and Balakrishnan 2010), (Cowan and Jonard 2004).

The micro-level study, in contrast, focuses on modeling each user as an agent making local adoption decisions based on the neighbors' adoptions of some item. The diffusion models developed at the micro-level require complete knowledge of user-user relationships, and the knowledge of influence probability between two neighboring users. Examples of these models include *Linear Threshold model* (Granovetter 1978), (Chen, Yuan, and Zhang 2010) and *Independent Cascade model* (Goldenberg, Libai, and Muller 2001), (Kempe, Kleinberg, and Tardos 2003), (Saito et al. 2010), (Gomez Rodriguez, Leskovec, and Krause 2010), the stochastic model for opinion formation of Sznajd (Dietrich and Stauffer 2002), the model on the impact of the network characteristics on innovation diffusion (Liu, Madhavan, and Sudharshan 2005) (a variety of agent-based models can be found in the comprehensive review (Iribarren and Moro 2011)). These models are however not applicable in situations where the detailed network structure is not given or when the influence probabilities are missing due to privacy or data non-availability reasons.

In this research, we focus on macro-level diffusion models to minimize the amount of data required for modeling. As the macro-level diffusion models do not require detailed

knowledge about network structure, the amount of information required for modeling is small and the number of parameters to be learnt are very few. Our research specifically addresses the relationship between social network properties and diffusion process. We incorporate the skewed degree distribution of social networks into macro-level diffusion models so that the latter can better explain the diffusion dynamics among the social network nodes.

1.2 Research Objectives and Contributions

This study aims to develop macro-level diffusion models that incorporate specific network topology characteristics. In this paper, we consider only the degree distribution feature of the network and the internal influence due to neighboring nodes. The investigated degree distributions include the *power-law* and *exponential distribution*. In the former, the probability $P(k)$ of having degree k is proportional to a constant power of the degree whereas in the latter, $P(k)$ is proportional to an exponent of the degree. In notations,

1. Power-law distribution:

$$P(k) \sim k^{-\alpha} \text{ for some constant } \alpha. \quad (1)$$

2. Exponential distribution:

$$P(k) \sim e^{-k/\lambda} \text{ for some constant } \lambda. \quad (2)$$

The novel contributions accrued include:

- We proposed two new models, *Scale-free network Linear Influence Model* (SLIM) and *Exponential network Linear Influence Model* (ELIM) for diffusion in Scale-Free (SF) and Exponential networks, which follow power-law and exponential degree distributions respectively. We show that the mathematical structures of the two diffusion models match the well known Bass model when assuming the node degree distribution is *static* during diffusion.
- Through synthetic diffusion data, we observe that the static degree distribution assumption does not hold as higher degree nodes are more likely to become adopters and thus making the node degree distribution more and more skewed in the later stages of diffusion.
- To circumvent the fitting errors of SLIM and ELIM due to evolving node degree distribution, we propose a multi-stage network linear influence model (MLIM) which adapts the degree distribution parameter in different stages of diffusion. MLIM is shown to produce smaller fitting errors in our experiments.

1.3 Outline of Paper

The outline of this paper is as follows. In Section 2, we briefly review related works in macro models. In Section 3, we establish a generic framework for deriving *macro* models. From this framework, the Bass Model can be retrieved as a special case. In order to incorporate the network topology, we further develop a probabilistic approach which will be applied to derive two new models, SLIM and ELIM, in Section 4. Since these two models are restricted to the case of *static* degree distribution among non-adopters, we propose a

multi-stage model, which adapts to dynamic degree distribution, in Section 5. For the evaluation of proposed models, the experiments were conducted and their results are presented in Section 6. Finally, we conclude our paper in Section 7.

2 Related Work

It is well known (e.g. (Newman 2003)) that most social networks exhibit small world characteristics such as short average path length, the cliquishness, the tendency of containing cliques or near-cliques, and the power-law degree distribution. The impact of these topological features on diffusion has been studied more and more lately. The works by Cowan (Cowan and Jonard 2004) and Schilling (Schilling and Phelps 2007) investigated the effect of the first two features, short average path length and the cliquishness, in a knowledge diffusion system. These works proposed that networks with high clustering and short average path lengths facilitate much more complete diffusion. Shaikh et al. developed a few macro diffusion models considering the effect of non-uniform internal influence throughout diffusion, multiple influence by multiple adopters and small world network property (Shaikh, Rangaswamy, and Balakrishnan 2010). Their proposed models' performance were empirically compared with BM and some other models using the Mean Square Error (MSE).

These works however do not consider the *degree distribution* property of social networks, which is also an essential feature. In the case of infectious diseases, the degree distribution was proved to have significant effect on their diffusion. By investigating power-law degree distributions, (Boguñá, Pastor-Satorras, and Vespignani 2003) proved that certain kind of these distributions can help disease diffuse widely. For information diffusion, the work by Nekovee (Nekovee et al. 2008) proposed a stochastic model for rumor spreading and examined the threshold behavior and dynamics of the model on different kinds of networks (random graphs, scale-free networks). This work used the mean-field approach which is similar to ours. However, there are two points that are essentially different from our work. First of all, there was no consideration related to the changing topology among non-adopters (termed *ignorant nodes* in the paper). Secondly, their proposed model also depends on the so-called *degree correlation function*, which may require more parameter estimations than our model.

3 Proposed Probabilistic Framework

3.1 General diffusion equation

We denote N as the whole population. For each time t , let $f(t)$ denote the proportion of new adopters at exactly that time and $F(t)$ denote the *cumulative* proportion of adopters up to time t . We denote the event adoption at t as a_t , adoption before t as A_t , no adoption before t as \bar{A}_t and $P(\text{adoption at } t | \text{no adoption before } t)$ as $P(a_t | \bar{A}_t)$ for brevity. Using these notations we define the general framework as follows.

Recall that the function $f(t)$ is the derivative of the func-

tion $F(t)$,

$$f(t) = \frac{dF}{dt}(t) \quad (3)$$

On the other hand, $f(t)$ is also given by the following conditional probability,

$$\begin{aligned} f(t) &= P(a_t|\bar{A}_t)P(\bar{A}_t) \\ &= P(a_t|\bar{A}_t)(1 - P(A_t)) \\ &= P(a_t|\bar{A}_t)(1 - F(t)) \end{aligned} \quad (4)$$

Then by equating the R.H.S of the two Equations (3) and (4), we obtain the following Ordinary Differential Equation (ODE) for the unknown function $F(t)$

$$\frac{dF}{dt}(t) = P(a_t|\bar{A}_t)(1 - F(t)) \quad (5)$$

Equation (5) provides a general equation for formulating various macro diffusion models. The most important component which helps to differentiate between various model is the adoption probability $P(a_t|\bar{A}_t)$.

3.2 Formulation of adoption probability

It is well-known that diffusion is under two kinds of influence, *external* and *internal*. The former refers to the influence from mass media on members of the market and the latter, also known as *word-of-mouth* (WOM) effect in literature, refers to the influence that adopted users exert on potential adopters. Upon exerting on a user, these two influences will create two different adoption probabilities, which are denoted as $P_{ext}(a_t|\bar{A}_t)$ and $P_{int}(a_t|\bar{A}_t)$ respectively.

It is noteworthy that these two influences may have *different weights*. For instance, a user may have 25% and 75% chance to be under external and internal influence respectively. Hence, it is reasonable to assign weights w_e and $1 - w_e$ for the external and internal influence respectively. This argument leads to the following formula for adoption probability

$$P(a_t|\bar{A}_t) = w_e \cdot P_{ext}(a_t|\bar{A}_t) + (1 - w_e) \cdot P_{int}(a_t|\bar{A}_t) \quad (6)$$

It is reasonable to assume that the component adoption probability $P_{ext}(a_t|\bar{A}_t)$ created by external influence is a constant p_e since this influence is stable. Therefore, the above formula is rewritten as:

$$P(a_t|\bar{A}_t) = w_e \cdot p_e + (1 - w_e) \cdot P_{int}(a_t|\bar{A}_t) \quad (7)$$

By instantiating different $P_{int}(a_t|\bar{A}_t)$, we can derive different diffusion models. For Bass Model (BM) with *full connectivity* assumption, $P_{int}(a_t|\bar{A}_t) \sim F(t)$, which can be seen in the subsequent section. However, when other network topologies are assumed, the computation of this component is much more complicated. Hence, it will be done separately in Section 3.4 after BM is reviewed.

3.3 Bass Model

Derivation of BM In (Bass 1969), Frank Bass assumed that the external influence is simply a constant p and the

internal influence is proportional to $F(t)$. Hence, the total adoption probability has the following simple form

$$P(a_t|\bar{A}_t) = p + q \cdot F(t) \quad (8)$$

By comparing with (7), the terms p and $q \cdot F(t)$ can be identified with the terms $w_e \cdot p_e$ and $(1 - w_e) \cdot P_{int}(a_t|\bar{A}_t)$ respectively.

Substituting (8) into the equation (5), we get

$$\frac{dF}{dt}(t) = (p + q \cdot F(t))(1 - F(t)) \quad (9)$$

Equation (9) is a first-order ODE Riccati equation. Hence, by (Enns and McGuire 2007), it can be solved analytically to obtain the closed form solution,

$$F_{BM}(t) = \frac{e^{(p+q)t} - 1}{e^{(p+q)t} + \frac{q}{p}}$$

and

$$f_{BM}(t) = F'_{BM}(t) = \frac{(p+q)^2}{p} \cdot \frac{e^{(p+q)t}}{\left(\frac{q}{p} + e^{(p+q)t}\right)^2} \quad (10)$$

Discussion on Bass Model The simple form of BM leads to a closed form solution in which the parameters can be estimated conveniently using any standard least square (LS) or maximum likelihood (ML) procedures. However, it has some significant restrictions due to the following issues.

1. The obvious violation of *full connectivity* assumption in practical networks with specific degree distributions.
2. The *non-uniform* WOM effect in the population, which requires the justification of the form $q \cdot F(t)$ of internal influence. Empirical data have shown that different persons in the population receive different influence. In fact, the influence that each person receives depend on his *network exposure*. One of the measures for individuals' network exposure is his *degree* (or *out degree*), the number of his friends (in undirected networks) or his followees (in directed networks). This again motivates us to incorporate degree distribution into our study.
3. The possibility of violating the constraint $P(a_t|\bar{A}_t) \leq 1$ due to its linear form. Thus, BM may overfit the empirical data. At this point, the generic form (7) proves its first advantage over BM, it can guarantee the probabilistic constraint. In fact, it can be seen in (7) that the convex combination of two quantities $P_{ext}(a_t|\bar{A}_t)$, $P_{int}(a_t|\bar{A}_t)$, both less than 1, guarantees the constraint.

Now that the final issue is already addressed by the generic form (7), from now on we will focus on the remaining ones, which depend on computing the component $P_{int}(a_t|\bar{A}_t)$ based on topology of the underlying network.

3.4 Modelling Internal Influence

Let us consider a specific social network with a given topology. Since the amount of influence which a user receives depends on his degree, the probability of adoption due to internal influence will vary along with users' degree. Hence

$P_{int}(a_t|\overline{A}_t)$ should be replaced by its *expected value*, which is taken over the degree distribution of the whole network. However, for the sake of brevity, the notation $P_{int}(a_t|\overline{A}_t)$ is still used in place of its expected value.

Let the random variables K denote *degree* of a person in the network and J be the number of adopters among its neighbours. J then follows a binomial distribution,

$$P(J = j|K = k) = \binom{k}{j} F_t^j (1 - F_t)^{k-j}. \quad (11)$$

here F_t is used in place of $F(t)$ for brevity.

By the law of total probability, we have

$$P_{int}(a_t|\overline{A}_t) = \sum_{k=1}^{N-1} \sum_{j=0}^k P(a_t, k, j|\overline{A}_t, a_t^{int}) \quad (12)$$

$$= \sum_{k=1}^{N-1} \sum_{j=0}^k P(a_t|j, \overline{A}_t, a_t^{int}) P(j|k) P(k) \quad (13)$$

Substituting into Equation (13) using the R.H.S of the equation (11), we obtain,

$$P_{int}(a_t|\overline{A}_t) = \sum_{k=1}^{N-1} P(k) \sum_{j=0}^k \left[\binom{k}{j} (F_t)^j (1 - F_t)^{k-j} P(a_t|j, \overline{A}_t) \right] \quad (14)$$

Assuming the degree distribution $P(K = k)$ is known, we can obtain the formulae for $P_{int}(a_t|\overline{A}_t)$ once a specific form is defined for the term $P_{int}(a_t|J = j, \overline{A}_t)$ in (14). This term is nothing other than the *probability of adoption due to j adopted neighbors*. In this work, we assume the simplest form, $P_{int}(a_t|J = j, \overline{A}_t) = c \cdot j$ and then derive the essential component $P_{int}(a_t|\overline{A}_t)$ as a function of F_t . For that derivation, the following technical proposition is needed.

Lemma 1. *Assume that the probability of adoption due to j adopted neighbors is linear i.e.*

$$P_{int}(a_t|J = j, \overline{A}_t) = c \cdot j \quad (15)$$

Then we get the following simplified equation

$$\sum_{j=0}^k \binom{k}{j} (F_t)^j (1 - F_t)^{k-j} P(a_t|j, \overline{A}_t, a_t^{int}) = c \cdot k \cdot F_t \quad (16)$$

Proof. Using the linear form (15), we have

$$\begin{aligned} & \sum_{j=0}^k \left[\binom{k}{j} (F_t)^j (1 - F_t)^{k-j} P_{int}(a_t|j, \overline{A}_t) \right] \\ &= \sum_{j=0}^k \left[\binom{k}{j} (F_t)^j (1 - F_t)^{k-j} c \cdot j \right] \\ &= c(1 - F_t)^k \sum_{j=0}^k \binom{k}{j} j \left(\frac{F_t}{1 - F_t} \right)^j \end{aligned}$$

Let $x = \frac{F_t}{1 - F_t}$, the final expression becomes

$$c(1 - F_t)^k \sum_{j=0}^k \binom{k}{j} j x^j$$

By standard manipulations, we have

$$\begin{aligned} & c(1 - F_t)^k \sum_{j=0}^k \binom{k}{j} j x^j \\ &= c(1 - F_t)^k x k (1 + x)^{k-1} \end{aligned}$$

Upon replacing $x = \frac{F_t}{1 - F_t}$, the final term will become $c \cdot k \cdot F_t$, we obtain the required Equation (16). \square

With this simplification, now we only need to substitute each of the two specific degree distributions (1), (2) in order to derive the two corresponding models in the subsequent sections.

4 Proposed Models for Networks with Specific Degree Distributions

4.1 Scale-free Network Linear Influence Model (SLIM)

By (Newman 2003) the precise form for power-law degree distribution is $P(k) = \frac{1}{\zeta(\alpha)} \cdot k^{-\alpha}$, where $\zeta(\cdot)$ is the Riemann's zeta function (c.f (Titchmarsh and Heath-Brown 1986)). Employing the above proposition and substituting this form, the probability of adoption due to internal influence is rewritten as

$$P_{int}(a_t|\overline{A}_t) = \frac{1}{\zeta(\alpha)} \sum_{k=1}^{N-1} k^{-\alpha} c \cdot k \cdot F_t \quad (17)$$

$$= \frac{c}{\zeta(\alpha)} F_t \sum_{k=1}^{N-1} k^{1-\alpha} \quad (18)$$

Noticing that the final sum on the RHS is $\zeta(\alpha - 1)$, we finally finish the derivation of internal influence as below.

$$P_{int}(a_t|\overline{A}_t) = \frac{\zeta(\alpha - 1)}{\zeta(\alpha)} \cdot c \cdot F_t \quad (19)$$

$$= IC(c, \alpha) \cdot F_t \quad (20)$$

in which $IC(c, \alpha) = \frac{\zeta(\alpha - 1)}{\zeta(\alpha)} \cdot c$, the coefficient of internal influence, depends on c and α .

Now that the adoption probability due to internal influence has the simple form given by (20), the total adoption probability will be

$$P(a_t|\overline{A}_t) = w_e \cdot p_e + (1 - w_e) \cdot IC(c, \alpha) \cdot F_t \quad (21)$$

Substituting this into the general ODE (5) in the framework, we obtain the following ODE

$$\frac{dF}{dt} = [w_e \cdot p_e + (1 - w_e) \cdot IC(c, \alpha) \cdot F(t)] (1 - F(t)) \quad (22)$$

By letting $p = w_e \cdot p_e$ and $q = (1 - w_e) \cdot IC(c, \alpha)$, we arrive at the ODE of BM. Thus, it is straightforward to derive the following main result, the SLIM model.

Proposition 1. Consider the diffusion in an SF network with power α . Assume that the influence which a node receives from its adopted neighbors is linear, i.e., $P_{int}(a_t|J = j, \bar{A}_t) = c \cdot j$ where c is some constant. Under this assumption, the cumulative adopter fraction $F(t)$ can be computed by the formula

$$F_{SLIM}(t) = \frac{e^{(p+q_{SLIM})t} - 1}{e^{(p+q_{SLIM})t} + \frac{q_{SLIM}}{p}} \quad (23)$$

where $p = w_e \cdot p_e$, and

$$q_{SLIM} = (1 - w_e) \cdot IC(c, \alpha) = (1 - w_e) \frac{\zeta(\alpha - 1)}{\zeta(\alpha)} \cdot c \quad (24)$$

4.2 Exponential Network Linear Influence Model (ELIM)

For this model, the internal influence due to j adopters in one's neighbors, $P_{int}(a_t|\hat{A}_t, J = j)$ is still $c \cdot j$. However, the underlying network now has an exponential degree distribution. Hence, on computing the adoption probability due to internal influence, instead of the power-law form, we substitute the exponential form $P(k) = [1 - e^{(-1/\lambda)}] e^{(-k/\lambda)}$ and obtain

$$P_{int}(a_t|\bar{A}_t) = \sum_{k=1}^{N-1} [1 - e^{(-1/\lambda)}] e^{(-k/\lambda)} \times \sum_{j=0}^k \left[\binom{k}{j} (F_t)^j (1 - F_t)^{k-j} c \cdot j \right] \quad (25)$$

$$P_{int}(a_t|\bar{A}_t) = c \cdot F_t \cdot \sum_{k=1}^{N-1} k [1 - e^{(-1/\lambda)}] e^{(-k/\lambda)}$$

Let $x = e^{(-1/\lambda)}$ then the final sum is rewritten simply as

$$c \cdot F_t \cdot (1 - x) \sum_{k=1}^{N-1} k \cdot x^k$$

After some standard manipulations, we obtain

$$\begin{aligned} P_{int}(a_t|\bar{A}_t) &= c \cdot F_t \cdot (1 - x) \sum_{k=1}^{N-1} k \cdot x^k \\ &= c \cdot \frac{e^{(-1/\lambda)}}{1 - e^{(-1/\lambda)}} \cdot F_t \end{aligned} \quad (26)$$

From this, we can proceed quite the same as SLIM model. In fact, by letting $p = w_e \cdot p_e$ and $q = (1 - w_e) \cdot c \cdot \frac{e^{(-1/\lambda)}}{1 - e^{(-1/\lambda)}}$, we also arrive at a Riccati equation and obtain the following main result for ELIM.

Proposition 2. Consider a static network whose exponential degree distribution is characterized by parameter λ . The parameters p_e , w_e , and c are defined as in proposition 1.

Then the cumulative adoption fraction $F(t)$ in this network can be computed by the formula

$$F_{ELIM}(t) = \frac{e^{(p+q_{ELIM})t} - 1}{e^{(p+q_{ELIM})t} + \frac{q_{ELIM}}{p}} \quad (27)$$

where $p = w_e \cdot p_e$ and

$$q_{ELIM} = (1 - w_e) \cdot c \cdot \frac{e^{(-1/\lambda)}}{1 - e^{(-1/\lambda)}} \quad (28)$$

5 Multi-Stage Macro-level Diffusion

5.1 Dynamic Degree Distribution

In defining the SLIM and ELIM models, we assume that the degree distribution under consideration is static throughout diffusion. This assumption unfortunately may not hold. As diffusion unfolds, the degree distribution of the network consisting of the remaining non-adopters actually changes.

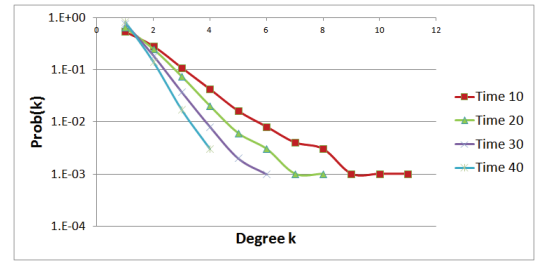


Figure 1: The curves (plotted in linear-log scale) depict changing degree distributions (among non-adopters) over time. This was recorded from diffusion on a SF graph $G = (V, E)$ with $|V| = 27289$, $|E| = 27031$, $\alpha = 3$, $p_e = 0.1$, $w_e = 0.2$, and $c = 0.108$

As shown in Figure 1, we examine the degree distribution of a scale free network consisting of remaining non-adopters at different time steps (=10,20,30 and 40) with adoptions synthetically generated. We observe that high degree nodes are more likely to adopt than low degree ones. Hence, the proportion of high/low degree nodes decreases/increases with increasing time steps. This makes the degree distribution among non-adopters more biased towards low degrees. Figure 1 also reveals that the degree distributions at the later time points may no longer follow power law. Instead, they may be more similar to exponential distributions (since the curves are nearly straight lines in log-linear scale).

Due to this dynamics of degree distribution, SLIM and ELIM which assume static degree distribution, will not do well in fitting the observed diffusion data particularly in the later time steps. To overcome this shortcoming, we propose to use a *multi-stage model*.

5.2 Multi-Stage Linear Influence Model (MLIM)

The main idea of a multi-stage model is to divide the whole diffusion period into a series of stages. In the early stages, which still have degree distribution following power law, we may fit the observed data using SLIM. For the later stages, we may fit the observed diffusion data using ELIM.

Assume that the diffusion process is divided into n stages, $[t_1 = 0, t_2), [t_2, t_3), \dots, [t_n, t_{n+1} = t_{max}]$. Now, the parameters w_e, p_e and c are kept constant but the degree distribution function and parameter are allowed to vary in different stages. The exact degree distribution parameter used depends on the choice of distribution function, i.e., α for power law and λ for exponential.

For each i -th stage, from time points t_i to t_{i+1} , we use the following notations:

- $F_i(t), f_i(t)$: the functions for cumulative and non-cumulative adoption proportions, respectively.
- α_i, λ_i : the parameter of the power-law or exponential degree distribution. In our experiments, we selected the one that gave the smallest fitting error.
- q_i : refers to q_{SLIM} or q_{ELIM} depending on the choice of degree distribution function (see Equations (24) and (28)).

With these notations, we can apply exactly the argument in previous section for stage i and solve for the closed form of functions $F_i(t)$ and $f_i(t)$. The only difference is that the initial condition $F(0) = 0$ (for the Riccati ODE) is now replaced by $F_i(t_i) = F_i^{start}$, the cumulative adoption proportion at t_i observed from data.

Proposition 3. *In the i -th stage, the closed forms of functions $F_i(t), f_i(t)$ are given by*

$$F_i(t) = F_i^{start} \cdot \frac{p_e \cdot w_e + q_i + q_i \cdot \Delta_t}{p_e \cdot w_e + q_i + q_i \cdot F_i^{start} \cdot \Delta_t} \quad (29)$$

and

$$f_i(t) = f_i^{start} \cdot (p_e \cdot w_e + q_i)^2 \cdot \frac{1 + \Delta_t}{\{p_e \cdot w_e + q_i + q_i \cdot F_i^{start} \cdot \Delta_t\}^2} \quad (30)$$

where

$$\Delta_t = \exp\{(p_e \cdot w_e + q_i)(t - t_i)\} - 1$$

and f_i^{start} , the value of f_i at t_i , is computed from F_i^{start} using Equation (9), i.e.,

$$f_i^{start} = (p_e \cdot w_e + q_i \cdot F_i^{start}) \cdot (1 - F_i^{start}) .$$

6 Experiments

6.1 Synthetic Data

In the subsequent experiments, we evaluate our proposed models and compare them with the well known Bass Model in modeling diffusion of items within a social network. The models are divided into two groups: *one-stage models* and *multi-stage models*. For all multi-stage models, the default number of stages is $n = 5$. A summary of the parameters is provided in Table 1. Note that SLIM and ELIM can be derived from BM’s parameters as mentioned in Propositions 1 and 2, they are expected to share identical data fitting errors as BM. Hence we do not show their results in our experiments.

¹ $MaxDeg = 78$ is the maximum degree of graph G

	Models	Parameters
1-stage	BM (baseline)	p, q
	SLIM	p_e, w_e, c, α
	ELIM	p_e, w_e, c, λ
multi-stage	MLIM	$n, p_e, w_e, c, \{\alpha_i\}$ (or $\{\lambda_i\}$)

Table 1: Model Parameters.

Parameters	Default value
p_e	0.1
w_e	0.2
c	$5/MaxDeg^1$

Table 2: Default parameter setting.

6.2 Dataset Generation

To generate synthetic data, we used Algorithm 1 to simulate diffusion on a default scale free network $G = (V, E)$ with $\alpha=2.5, |V| = 28,172, |E| = 34,758$. With ground truth parameters p_e, w_e and c (values provided in Table 2), this algorithm generates the set of new adopters at each time step $t \in [1, tmax]$ using the linear influence assumption (see Equation (15)), computes the discrete version \hat{f}_t of the adoption proportion $f(t)$ and updates the set of non-adopters \mathcal{NA} .

6.3 Evaluation metrics

In all experiments, by regressing the function $f(t)$ with its discrete version \hat{f}_t , we learned the parameters p_e and α_i ’s (or λ_i ’s) assuming that c and w_e were already known. In this way, we avoid the non-identification problem, i.e. non-uniqueness of solutions, which happens when we learned all parameters simultaneously. After learning the parameters, we can evaluate the results based on two metrics: (a) *model-fitting error* which is the least square error (LSE) between model and synthetic data, and (b) *parameter error* which is the difference between the learned and the ground truth parameter.

6.4 Results with different external influence weights

We first evaluate the performance of different models when the external influence weight changes. We generated datasets with different w_e ’s from $\{0.05, 0.1, 0.2, 0.3, 0.4\}$. Other parameters were fixed with default values. For each w_e value, ten synthetic dataset instances were generated. We then applied the different models, i.e., BM and MLIM, to learn their parameters.

Figure 2(a) shows the average model-fitting error of the two models over ten dataset instances. We observe that the multi-stage model (MLIM) always outperformed the 1-stage model (BM), especially for small w_e ’s. This can be explained by larger contribution to diffusion from internal influence when w_e is small. Obviously, this influence is affected a lot by the dynamics of degree distribution among non-adopters, which is much better modeled by MLIM. Though not shown in the figure, MLIM also outperformed

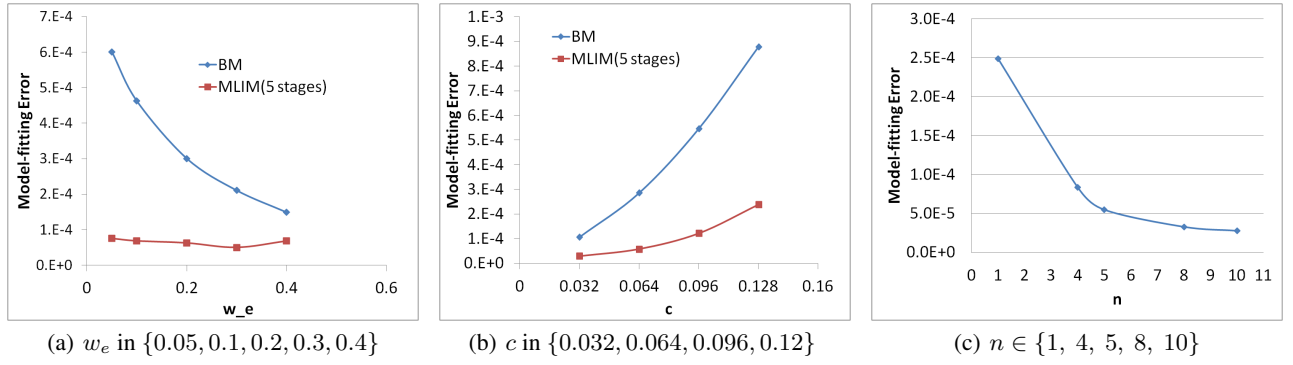


Figure 2: Model-fitting error under different parameter settings: varying w_e , c or n

Algorithm 1: GenerateSynData($G, p_e, w_e, c, tmax$)

```

     $G = (V, E)$  graph representing the SF network
Input :  $p_e, w_e, c$  the parameters of SLIM
           $tmax$  the final time of adoption
Output: vector  $\hat{f} = (\hat{f}_1, \dots, \hat{f}_{tmax})$  storing adoption
          proportion at different time points.
begin
    // At  $t=1$ , initialize randomly a
    // seed set  $\mathcal{S}$  of size  $\lfloor p_e \cdot w_e \cdot |V| \rfloor$ 
     $\hat{f}_1 \leftarrow p_e \cdot w_e$ ;
     $\mathcal{NA} = V \setminus \mathcal{S}$ ;
    // generate diffusion data for
     $t = 2 \rightarrow tmax$ 
    for  $t = 2$  to  $tmax$  do
        // update the adoption prob of
        // each non-adopter
        foreach  $x \in \mathcal{NA}$  do
             $j \leftarrow |\text{AdoptedNeighbors}(x)|$ ;
             $P_{int}(a_t | j) \leftarrow c \cdot j$ ;
             $\text{AdoptPr}(x) \leftarrow w_e \cdot p_e + (1 - w_e) \cdot P_{int}(a_t | j)$ ;
        end
        // determine new adopters using
        // updated adoption probs
         $\text{NewAdopters} \leftarrow \emptyset$ ;
        foreach  $x \in \mathcal{NA}$  do
            If  $\text{AdoptPr}(x) > \text{thres}$  then add  $x$  to the
            // set  $\text{NewAdopters}$ ;
        end
        // update diffusion data
         $\mathcal{NA} \leftarrow \mathcal{NA} \setminus \text{NewAdopters}$ ;
         $\hat{f}_t \leftarrow \frac{|\text{NewAdopters}|}{|V|}$ ;
    end
return  $\hat{f}$ 
end

```

BM in learning p_e . The average relative p_e error of MLIM and BM were 7.71% and 17.23% respectively.

6.5 Results with different linear influence coefficients

We next generated datasets with different coefficients c in $\{i * \frac{2.5}{\text{MaxDeg}}, i = 1, 4\} = \{0.032, 0.064, 0.096, 0.12\}$ using our default scale free graph. For each c , we also generated ten dataset instances.

Since increasing c also increases the internal influence, the effect of increasing c is expected to be similar to that of decreasing w_e . Indeed, this expectation is verified by the results in Figure 2(b). We observe that MLIM again outperforms BM for different c . MLIM could also learn the parameter p_e better than BM. The average relative p_e error of MLIM and BM were 3% and 12% respectively.

6.6 Results with different number of stages

For this experiment, n was varied so that we could evaluate the performance of MLIM over different number of stages. As shown in Figure 2(c), the more stages, the better MLIM performs. Moreover, the relative performance difference between one-stage model (BM) to multistage model can be quite substantial. These observations are reasonable given that more parameters can be learnt at different stages to fit the observed data when n is large.

6.7 Experiments on Real Dataset

We conducted experiments on a dataset from GoodReads (<http://www.goodreads.com/>), a popular website for recommending and sharing books. GoodReads users have follow relationships among one another. In this experiment, we consider a user to have adopted a book when she wrote review about the book. The GoodReads user network was collected from Jan 21, 2011 to March 15, 2011. It consists of nearly 86K users and 159,442 follow links. We selected 20 books which have reviews that span over at least 30 months so as to allow us to observe their diffusion in the user network. The chosen books are among the most popular ones (e.g. Harry Potter 7, ID=136251; Breaking Dawn, ID=1162543) so that their diffusion processes are guaranteed to be long and proper enough for our study.

BookId	BM	MLIM	$\frac{MLIM}{BM}$
1162543	0.001	0.0004	0.355
14866	0.011	0.007	0.623
1609451	0.009	0.006	0.736
297673	0.011	0.008	0.742
25460	0.006	0.005	0.771
428263	0.0002	0.0002	0.772
693208	0.005	0.004	0.782
248484	0.006	0.005	0.792
2213661	0.007	0.005	0.792
128029	0.005	0.004	0.798
2248573	0.003	0.003	0.807
1656001	0.002	0.002	0.813
1582996	0.002	0.002	0.821
345627	0.006	0.005	0.884
136251	0.0004	0.0003	0.918
2767052	0.002	0.002	0.976
30183	0.002	0.002	0.989
3236307	0.003	0.003	1.022
2120932	0.003	0.003	1.075
3777732	0.002	0.003	1.176
Avg	0.004	0.003	0.834

Table 3: The second and third column show the fitting error of corresponding models. The last column shows the ratio of the two errors.

For this experiment, there were no ground truth parameters. We therefore had to learn all the parameters of the models. The fitting error results for the 20 selected books are shown in Table 3. The results show that MLIM performs better than BM for most books.

7 Conclusions

In this work, we proposed several macro-diffusion models which incorporate the relationship between diffusion and degree distribution of the underlying network. We developed two models (SLIM and ELIM) for *static* degree distribution and one model (MLIM) for *dynamic* degree distribution among non-adopters which can be observed in the diffusion process. Experiments on synthetic data show that the performance of our proposed models is promising for fitting purpose as well as for recovering parameters. The future works include extending the models to predict diffusion and developing a rigorous way to determine proper degree distribution (among non-adopters) for each stage in diffusion process.

Acknowledgement

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

References

Bass, F. M. 1969. A new product growth for model consumer durables. *Management Science* 15(5):215–227.

Boguñá, M.; Pastor-Satorras, R.; and Vespignani, A. 2003. Absence of epidemic threshold in scale-free networks with degree correlations. *Phys. Rev. Lett.* 90:028701.

Chen, W.; Yuan, Y.; and Zhang, L. 2010. Scalable influence maximization in social networks under the linear threshold model. In *Proceedings of the 2010 ICDM*, 88–97. IEEE Computer Society.

Cowan, R., and Jonard, N. 2004. Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control* 28(8):1557 – 1575.

Dietrich, and Stauffer. 2002. Sociophysics: the sznajd model and its applications. *Computer Physics Communications* 146(1):93 – 98.

Enns, R. H., and McGuire, G. C. 2007. Nonlinear ode models. In *Computer Algebra Recipes*. Springer New York. 149–206.

Goldenberg, J.; Libai, B.; and Muller, E. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters* 211–223.

Gomez Rodriguez, M.; Leskovec, J.; and Krause, A. 2010. Inferring networks of diffusion and influence. *KDD '10*, 1019–1028. ACM.

Granovetter, M. 1978. Threshold models of collective behavior. *American Journal of Sociology* 83(5):1420–1443.

Iribarren, J. L., and Moro, E. 2011. Affinity paths and information diffusion in social networks. *CoRR* abs/1105.3316.

Kempe, D.; Kleinberg, J.; and Tardos, E. 2003. Maximizing the spread of influence through a social network. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137–146.

Liu, B. S.; Madhavan, R.; and Sudharshan, D. 2005. DiffuNET: The impact of network structure on diffusion of innovation. *European Journal of Innovation Management* 8(2):240–262.

Meade, N., and Islam, T. 2006. Modelling and forecasting the diffusion of innovation - a 25-year review. *International Journal of Forecasting* 22.

Nekovee, M.; Moreno, Y.; Bianconi, G.; and Marsili, M. 2008. Theory of rumour spreading in complex social networks. *CoRR* abs/0807.1458.

Newman, M. E. J. 2003. The Structure and Function of Complex Networks. *SIAM Review* 45(2):167–256.

Saito, K.; Kimura, M.; Ohara, K.; and Motoda, H. 2010. Behavioral analyses of information diffusion models by observed data of social network. In *SBP*, 149–158.

Schilling, M. A., and Phelps, C. C. 2007. Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Management Science* 53(7):1113–1126.

Shaikh, N. I.; Rangaswamy, A.; and Balakrishnan, A. 2010. Modeling the diffusion of innovations using small-world networks. Working paper, Penn State University.

Titchmarsh, E., and Heath-Brown, D. 1986. *The theory of the Riemann zeta-function*. Clarendon Press.