10-2011

A survey of information diffusion models and relevant problems

Minh Duc LUU
Singapore Management University, victorluu@smu.edu.sg

Tuan Anh HOANG
Singapore Management University, tahoang@smu.edu.sg

Ee-Peng LIM
Singapore Management University, eplim@smu.edu.sg

A Survey of Information Diffusion Models and Relevant Problems

Duc M. Luu*, Hoang Tuan Anh† and Ee-Peng Lim‡
Singapore Management University, Singapore
Emails: * mdluu.2011@phdis.smu.edu.sg
† tahoang.2011@phdis.smu.edu.sg
‡ eplim@smu.edu.sg

I. Introduction

There have been tremendous interest in diffusion of innovations or information in a social system. Nowadays, social networks (offline as well as online) are considered as important medium for diffusion and large amount of research has been conducted to understand the dynamics of diffusion in social networks. In this work, we review some of the models proposed for diffusion in social networks. We also highlight the major features of these models and summarize the main results obtained so far. These findings aim at providing a clear, systematic view on existing models. Moreover, we strive to show the connections among these models. Finally, our survey also helps to review what have been achieved in this area and what may need to be improved.

We divide the surveyed models into two categories: *non-network* and *network* diffusion models. The former refers to user communities without any knowledge about the user relationship network and the latter is more applicable to the social networks where user relationships network are given (e.g. Facebook, blog networks). We first give brief reviews of the basic diffusion models. We then describe the applications as well as extensions of these models. The surveyed applications include *Influence Maximization* and *Contamination Minimization*. The extensions are the asynchronous models which incorporate *time delay* factor into the basic models. Finally, a concise table summarizing key features of the models is provided in Table I at the end of this survey.

II. Non-network diffusion models

In the literature of diffusion theory, the well known Bass Model (BM) [1] has been studied extensively since its introduction in the 1960s. This model assumes that potential adopters of an innovation are influenced by external influence (mass media) and internal influence (word of mouth). The former is represented by a constant P and the latter is proportional (by a constant Q) to the cumulative number of adopters, which depends on time. Under these assumptions, Bass formulated the following ordinary differential equation (ODE) for the cumulative proportion of adopters $F(t)$ (as a function of time).

$$\frac{dF}{dt} = (P + Q \cdot F(t))\,(1 - F(t)) \tag{1}$$

This ODE has an analytical solution, which allows Bass model to forecast the adoption rate as well as peak adoption time. This success of BM sparked considerable research and further extensions. The extensions include the Nonuniform influence (NUI) model proposed by Easingwood et al., the Flexible logistic growth (FLOG) model by Bewley et al. A comprehensive review about these models can be found in [2]. Each of these relaxed some assumption underlying the Bass model. For instance, the models NUI and FLOG do not fix the coefficient Q of internal influence, instead they allow it to vary systematically as a function of time. With different assumptions, each extension is applicable to diffusion for certain types of products.

III. Basic network diffusion models

One of the characteristics of the Bass model and its extensions is that they are *macroscopic* models. The parameters P and Q are determined at the aggregate level (over all users). Hence a natural question is how these parameters can be interpreted at the *microscopic*(individual) level. This question was answered in [3], which considered the external and internal influences at microscopic level. The former was assigned a constant-value probability p. For the latter, the case of homogeneous market was first considered so that q is also a constant and then this assumption was relaxed in the case of heterogeneous market.

In the case of homogeneous market, the probability of adoption $PA(t)$ at time t for a potential adopter x was given by

$$PA(x,t) = 1 - (1-p)(1-q)^{Y(x,t)} \tag{2}$$

where $Y(x,t)$ is the number of adopters in the neighbours of the node x.

Using least square regression, the authors determined the relationship between p, q and P, Q. As predicted, Q is mainly generated by q whereas P is mainly effected by p, although q also has some significant negative effect on P. The last inverse relation was reasoned that the estimation of P was not as clean as that of Q e.g. there were more noises in the former than the latter.

Noting that in (2), the underlying assumption is that the influences from different adopters are *independent*. Now, to account for heterogeneity, the constant q can be replaced by a matrix $P = (p_{vw})$, where p_{vw} is the probability that adopter

v can successfully activate a potential adopter w. Using the independence assumption again, the factor $(1-q)^{Y(t)}$ now becomes $\prod_w(1-p_{vw})$. This leads to a model quite similar to the well known Independent Cascade(IC) model briefly described in section III-A.

Another popular approach is the Linear Threshold (LT) model, which reflects the *threshold* nature of many decisions (e.g. adoption decisions). The motivation for this model lies in the fact that many decisions are inherently costly (even risky), requiring investment of time and resources. This requires that the value of the relevant decision function should reach some *threshold*; for instance the decision of switching state is only made when the total influence achieves at least a node specific threshold θ_v. Moreover, the node specific thresholds are allowed to vary randomly (w.r.t some distribution e.g $U[0,1]$) to account for variations knowledge, preferences, and observational capabilities across the population.

A. Independent Cascade (IC) model

The classical IC model assumes that the diffusion probabilities p_{vw}'s are given and that the cascading actions among different parent nodes are independent from one another.
The diffusion process is discrete in time and proceeds from an initial set S until no more activations can be made. When a node v becomes active at time t, it has a *single* chance of activating each currently inactive child (neighbor) w with the probability of success p_{vw}. If v succeed, w becomes active at time $t+1$. Whether or not v succeeds, it cannot make any further attempts to activate w in subsequent rounds. If multiple parent nodes of w become active at time t, their influence are all performed at time step t to $t+1$.

B. Linear Threshold (LT) model

In LT model proposed by [4], the influence on a node w was assumed to be a function solely of *relative* size of the set $Y(w,t)$ of adopters in its neighbors. Then this assumption was extended in a more general way. That is the influence not only depends on the size of $Y(w,t)$ but also on attributes of each element of $Y(w,t)$. Specifically, each node w is influenced by each parent v according to a specified weight λ_{vw} (e.g. $\lambda_{vw} = \frac{1}{|Y(w,t)|+1}$) such that the total influence does not exceed 1, i.e

$$\sum_{v \in Y(w,t)} \lambda_{vw} \leq 1$$

Here the total influence is measured as the sum of all individual influences, hence the name of the model.
The diffusion process from a given initial active set S proceeds according to the following randomized rule. First, for any node w, a threshold θ_w is chosen uniformly at random from the interval $[0;1]$. At time-step t, an inactive node w is influenced by each of its active parent nodes v according to weight λ_{vw}. If the total influence is not less than θ_w, i.e. $\sum_{v \in Y(w,t)} \lambda_{vw} \geq \theta_w$, then w becomes active at time $t+1$. The process terminates if no more activations are possible.

C. Unified framework for likelihood functions

Saito et al. further proposed a unified framework for likelihood functions of both models IC and LT in the works [5], [6]. They stated that, during an observed time period, the likelihood function is described by the product of two factors. The first factor represents the probabilities that adopted nodes are activated at exactly their respective times and the second represents the probabilities that susceptible nodes (i.e. the children of adopted nodes) have not been activated. The reason for this statement is that during that time period, the only nodes involved are the new adopted ones and its child nodes which are not activated yet.

Using this framework, the explicit formulae of likelihood functions were established and then maximized by *Expectation Maximization* (EM) algorithms to learn the respective parameters. After learning the parameters, the models were applied to solving many following problems:

1) *Finding and ranking influential node* (e.g. in [7], [8]). In [7], the proposed method could predict the high ranked influential nodes much more accurately than the well studied conventional four heuristic methods (degree centrality, closeness centrality, betweenness centrality, PageRank). In [8], the influence degree of each node was proved to depend on node attributes and substantially different from that obtained assuming a uniform diffusion probability (on every link).
2) Predicting *information diffusion probabilities* and calculating *expected influence degree* (of a node) or *contamination degree* (of a graph) (in [5], [9], [8]). The two last quantities are essential for the solution of important problems including influence spread maximization and contamination minimization. These two relevant problems will be described below.

IV. Maximizing influence spread problem for IC and LT models

Once basic dynamics of information diffusion has been understood, it is natural to consider the applications of maximizing influence spread (e.g. of opinion, idea, innovation, etc.) or conversely, minimizing contamination (e.g. computer viruses, spam emails, etc.). The former was addressed in the work [10],[11], [12], and the latter in [9], [13].

A. Influence spread maximization

1) Problem formulation: Given a budget with can be used to create a seed node set S of size k, how should we choose S to trigger the largest cascade of influence? If we denote $\sigma(A)$ the function measuring influence spread created by an arbitrary node set A, then this is a discrete optimization problem where the objective function is σ.
An extension of this problem was proposed in [11], which deals with the case when negative opinions may emerge and propagate. For this extension, this work also established results similar to those in [10], which were presented below.

2) Theoretical results: When the function σ is estimated using IC or LT models, it is proven in [10] that, generally, the complexity of this problem is *NP hard*. However, by exploiting the monotonicity and *submodularity* of the objective function σ (for both models), we can use a greedy algorithm for it to get an approximate solution within 63% of optimal for these classes of models. Moreover, for special kinds of graphs such as directed acyclic graphs, a scalable algorithm (e.g. linear in time) for LT model was proposed by Chen et al. in [12]. These authors also proposed a heuristic algorithm in [14] which was proved to be easily scalable to large scale networks with millions of nodes and edges.

3) Experiments: For experiments, Kempe et al. used a collaboration graph in physics publication with 10748 nodes, and edges between about 53000 pairs of nodes. The models were realized in simple settings as below.

1) LT model: the weights of links were just multiplicity of edges. If nodes u, v have $c_{u,v}$ edges between them,(i.e. co-author in $c_{u,v}$ papers), and degrees d_u, d_v then the edge (u, v) has weight $c_{u,v}/d_v$ and the edge (v, u) has weight $c_{u,v}/d_u$.
2) IC model: all probabilities p_{vw} were assigned a uniform value p. Again, if u, v have $c_{u,v}$ edges between them, then u has a total probability of $1 - (1 - p)^{c_{u,v}}$ of activating v.

With these experimental setups, the results showed that, for both models, the greedy algorithm outperforms other heuristics which chose high-degree or high distance centrality nodes.

B. Contamination minimization

1) Problem formulation: In [9], Kimura et al. defined the *influence degree* $\delta(v, G)$ of a node v on graph G. It is the expected number of active nodes at the end of the random process of the LT model on G when there is only an initial active node v. Using this, they defined the *contamination degree* $c(G)$ of graph G as the average of influence degrees of all the nodes in G, that is,

$$c(G) = \frac{1}{|V|} \sum_{v \in V} \delta(v, G) \tag{3}$$

Finally, for $D \subset E$, the set of edges in G, they denoted $G(D)$ as the graph obtained from G by blocking edges in D.
Then the contamination minimization problem was defined in [9] as follows: given a positive integer $k < |E|$, find the optimal D^* with $|D^*| = k$ such that $c(G(D^*)) \leq c(G(D)),\ \forall D, |D| = k$.
The work [13] tackled the problem rather differently by identifying the *good blockers*, the nodes that block effectively the spread of influence. Also, they investigated the effectiveness of many kinds of structural measures (e.g. degree, betweenness and so on) as indicators of blocking ability of individual nodes. Moreover, the model it used is the IC model.

2) Proposed methods: For this problem, the authors in [9] also proposed a greedy algorithm for approximate solution. However, the implementation of this algorithm required a method for estimating contamination degree $c(\tilde{G})$ (or equivalently influence degree $\sigma(v, \tilde{G})$), for a given graph $\tilde{G}$. The method used for estimating contamination degree based on bond percolation process, which randomly designates each link of network G either "occupied" or "unoccupied" according to some probability distribution (refer to [15] for details). In [13], to find best blockers, Habiba et al. used exhaustive search method, which is computationally expensive. Hence, they only conducted limited experiments on their datasets.

3) Experiment: The experiments in [9] used two large real networks, blog and Wikipedia networks. The former had $12,047$ nodes and $79,920$ directed links, and the latter network had $9,481$ nodes and $245,044$ directed links. By comparing the proposed method with two heuristics, which were based on betweenness and out-degree, it was observed that the greedy algorithm outperformed the heuristics.

The datasets in [13] included a co-citation network (from $1967 - 2005$) in DBLP, the Enron email network and so on. The results obtained from comparing 17 investigated measures showed four measures that performed consistently well as blocker indicators (refer to [13] for details).

V. Asynchronous diffusion models

Both IC and LT models have parameters that need be specified in advance: diffusion probabilities for the former, and weights for the latter. This poses the problem of estimating these parameters from a set of information diffusion results. This problem was addressed by Saito et al. in [5] by maximizing the respective likelihood function. Hence it is crucial to construct a good likelihood function. Upon conducting this construction, one important factor that needs a special care is how to treat time delay in information diffusion. In the basic models, no time delay is considered, in other words, every action is uniformly delayed by one discrete time step. However, to do realistic analyses of information diffusion, it is necessary to be able to cope with *asynchronous* time delay. This is the motivation of the works [6], [16], which established asynchronous IC(AsIC) and LT(AsLT) models. Before delving into those models, it is worth to briefly introduce some major definitions.

A. Notions of time delay

There are two types of time delay considered: *link delay* and *node delay*. The former is associated with propagation delay and the latter is with action delay. Propagation delay is present in blog posting where a blogger u posts some article and it takes time before another blogger v reads the article (activated). Action delay can be seen in the case of email. If u sends an email to v, it is natural to assume that the email reaches v immediately, i.e. there is no delay in information diffusion from u to v. However, there is no guarantee that v will read that email as soon as he receives it, which leads to a delay in his action. Further, when v notices the mail, v may think to respond to it later. But before v responds, a new mail may arrive which needs a prompt response and v sends a mail immediately. We can think of this as an update of acting time

TABLE I
SUMMARY OF KEY FEATURES

	Non-network models	Network models
Features	Macroscope	Microscope
	Not consider topology of network	Consider topology of network
	Continuous time	Discrete time
Parameter estimation	Simple Maximum likelihood function	Complex Maximum likelihood function $\mathcal{L}$
	(Weighted) Least square error regression	EM type algorithm for regression (maximizing $\mathcal{L}$)

or an override of decision. In summary, node delay can go with either override or non-override, and link delay can only go with non-override.

B. Likelihood function and properties of asynchronous models

Saito et al. formulated explicitly likelihood functions (AsIC, AsLT) which incorporated these above notions by introducing a new parameter for time delay r_u. This parameter was chosen from an exponential distribution. Since these formulae are complicated, we do not provide it here but refer readers to [6]. Instead we focus on summarizing the following properties of these two models.

1) Expected influence degree: As discussed above, this quantity plays an important role in solving several important problems such as influence maximization and contamination minimization. For the purpose of obtaining this quantity only, it suffices to use basic models with no delay. It is because the expected influence degree obtained by basic models are the same as the one provided by asynchronous models after a substantially large time has passed. However, if we need to estimate expected influence degree at a specific time then the asynchronous models become very essential.
2) Behavioral analyses: Both [6] and [17] investigated the models AsIC and AsLT in terms of the sensitivity of the estimated parameters with respect to the topic of information. Saito et al. observed that regardless of difference in the models, the results were very similar.

C. Experiments

In [6] and [17], the experiments were implemented using four real datasets, which are all bidirectional connected networks. These included blog, Wikipedia, Enron email and co-authorship networks. The authors assumed the simplest case where $\lambda_{u,v} = q|Y(v,t)|^{-1}, \forall u \in Y(v,t)$ and $r_v = r$. With ground truth values $q = 0.9$ and $r \in \{2, 1/2\}$, their proposed method using EM algorithm proved to be effective. In the case of AsLT model, they also measured the influence degree and used it to rank the nodes. This ranking result was then compared with four heuristics widely used in social network analysis (degree centrality, closeness centrality, betweenness centrality, PageRank) and the AsIC model. Again the proposed method under AsLT model gave better results.

VI. CONCLUSION

In summary, the non-network diffusion models, whose representative is Bass model, are applicable at macroscopic level and can be used to predict the aggregated quantities such as adoption proportion or peak adoption time. The network diffusion models explore deeper to microscopic level and consider node (or edge) specific quantities such as the diffusion probabilities p_{vw}'s (in IC model) or the thresholds θ_v's (in LT model). Thanks to the works [5], [8] etc., these parameters can be estimated quite accurately. These estimates in turn facilitate finding solution for important applications such as *Influence Spread Maximization* and *Contamination Minimization*. Moreover, the network models can still be made more realistic by incorporating the time delay factor, which was conducted in [6], [16]. These recent works are promising and with appropriate investigation, we will likely to see more interesting results as well as applications in the future.

ACKNOWLEDGEMENT

This project was carried out at the Living Analytics Research Centre sponsored and supported by the Singapore National Research Foundation, Interactive & Digital Media Program Office, Media Development Authority.

REFERENCES

[1] F. M. Bass, "A new product growth for model consumer durables," *Management Science*, vol. 15, no. 5, pp. 215–227, 1969.
[2] V. Mahajan, E. Muller, and F. M. Bass, "New product diffusion models in marketing: A review and directions for research," *The Journal of Marketing*, vol. 54, no. 1, pp. 1–26, 1990.
[3] 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*, pp. 211–223, Aug. 2001.
[4] D. J. Watts, "A simple model of global cascades on random networks," *Proceedings of The National Academy of Sciences*, vol. 99, pp. 5766–5771, 2002.
[5] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *KES (3)*, 2008, pp. 67–75.
[6] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Generative models of information diffusion with asynchronous timedelay," *Journal of Machine Learning Research - Proceedings Track*, vol. 13, pp. 193–208, 2010.
[7] M. Kimura, K. Saito, R. Nakano, and H. Motoda, "Finding influential nodes in a social network from information diffusion data," in *Social Computing and Behavioral Modeling*. Springer US, 2009, ch. 18, pp. 1–8.
[8] K. Saito, K. Ohara, Y. Yamagishi, M. Kimura, and H. Motoda, "Learning diffusion probability based on node attributes in social networks," in *ISMIS*, 2011, pp. 153–162.

[9] M. Kimura, K. Saito, and H. Motoda, "Solving the contamination minimization problem on networks for the linear threshold model," in *PRICAI 2008: Trends in Artificial Intelligence*, 2008, vol. 5351, pp. 977–984.
[10] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," ser. KDD '03, 2003, pp. 137–146.
[11] W. Chen, A. Collins, R. Cummings, T. Ke, Z. Liu, D. Rincón, X. Sun, Y. Wang, W. Wei, and Y. Yuan, "Influence maximization in social networks when negative opinions may emerge and propagate," in *SDM*, 2011, pp. 379–390.
[12] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," ser. ICDM '10.
[13] H. Habiba, Y. Yu, T. Y. Berger-Wolf, and J. Saia, "Finding spread blockers in dynamic networks," ser. SNAKDD'08.
[14] 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 '10*, 2010, pp. 1029–1038.
[15] M. Kimura, K. Saito, R. Nakano, and H. Motoda, "Extracting influential nodes on a social network for information diffusion," *Data Mining and Knowledge Discovery Journal*, vol. 20, pp. 70–97, January 2010.
[16] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Selecting information diffusion models over social networks for behavioral analysis," in *ECML/PKDD (3)*, 2010, pp. 180–195.
[17] ——, "Behavioral analyses of information diffusion models by observed data of social network," in *SBP*, 2010, pp. 149–158.