Association for Information Systems
AIS Electronic Library (AISeL)

UK Academy for Information Systems Conference Proceedings 2010

UK Academy for Information Systems

Spring 3-23-2010

AN APPROACH TO APPROXIMATE DIFFUSION PROCESSES IN SOCIAL NETWORKS

Markus Helfert
Dublin City University, Markus.Helfert@computing.dcu.ie

Juan Yao
Dublin City University, jyao@computing.dcu.ie

AN APPROACH TO APPROXIMATE DIFFUSION PROCESSES IN SOCIAL NETWORKS

Markus Helfert, Juan Yao
School of Computing, Dublin City University, Dublin, Ireland
{Markus.Helfert, jyao}@computing.dcu.ie

Abstract

Social network analysis is concerned with the analysis of influence of an individual within a social network and how the influence diffuses through the network. It has been shown useful in business analytics. In this paper, we extend a nonlinear dynamical system that accurately models virus propagation in epidemiology to model information diffusion in social networks. Our approach can numerically calculate each node's probability to get activated given the initial active set. It provides an alternative way of estimating the number of nodes reached by the initial target set in the diffusion process. We validate our approach by comparing its predicting performance with diffusion simulations. Using the number of nodes reached in the diffusion process as an influence measure, our results show that the proposed method can provide a way of identifying nontrivial nodes as influencer.

1.0 Introduction

Business analytics that leverages customer data through profiling, segmentation and predictive modelling is widely adopted by many customer-driven companies. For instance, a customer retention program based on neural networks aims to predict customers' likelihood to churn. Typically, the predictive model is created using regular corporate attributes like billing history, usage behaviour and the set of products and services purchased. Yet many companies still face low customer response rates to marketing initiatives, coupled with increasing customer churn. Research (Gabbott and Hogg, 1994) indicates that customers rely on each other's judgment and experience when making purchasing and loyalty decisions. Domingos and Richardson (2001) argues that ignoring the network value of customers may lead to very suboptimal marketing decisions. Social network analysis provides an alternative view, where the individual customer characteristics are less important than their interactions. It is concerned with the analysis of the influence of an individual within a social network and how the influence propagates through the network. According to Doyle (2008) the value of analyzing customer networks is that the insight can help improving customer value management, churn measurement and up sell campaign performance.

A fundamental question is to measure the influence value in customer network and further identify influential customers. Intuitively, high connectivity in the network could be a factor as the commonly used centrality based heuristics in the sociology literature. However, a customer who is not widely connected may in fact have high influence value if one of the neighbours is highly connected (Domingos and Richardson, 2001). A customer's influence does not end with the immediate neighbours. Those neighbours may in turn influence their own neighbours and possibly lead to a cascade of influence. This is closely related to the diffusion processes in social networks – a phenomenon has been observed in many cases – the sudden widespread popularity of new products or services gained from word-of-mouth effect; the transmission of infected diseases or computer virus; the propagation of hot topics or rumours on blogs. Clearly, the number of customers reached by that customer in the diffusion process could be an important factor. The question is how to evaluate the number of customers reached? It's an open question to compute this quantity exactly by an efficient method, but very good estimates can be obtained by simulating the random diffusion process thousands of times to reach equilibrium (Kempe, Kleinberg and Tardos, 2003). However, according to Estevez, Vera and Saito (2007) this approach has a heavy computation load.

Models for the processes by which ideas and influence propagate through a social network have been studied in a number of domains. Most of the previous research has been done in the context of epidemiology and the spread of diseases over the network. In this paper, motivated by the great potential of social network marketing, we extend a nonlinear dynamical system (NLDS) that accurately models virus propagation in any arbitrary network (Chakrabarti *et al.*, 2008) from epidemiology. The epidemic model provides our work with the following contributions: (1) our approach can numerically calculate each customer's probability to get activated given the initial target set. By examining the probability values evolve over time step we can have a dynamic view of how influence spreads through customer network. (2) We provide an alternative way of estimating the number of nodes can be reached by an initial target set at the end of the diffusion process. Computational experiments results show that the sum of all probability values in the network gives an approximate estimation to the number of nodes reached by the target set. (3) Using the number of nodes reached in the diffusion process as an influence measure, our results show that our proposed method can provide a way of identifying nontrivial nodes as influencer.

The rest of the paper is organized as follows. In section 2 we give a general review of previous theoretical diffusion models and different approaches of measuring a node's influence in social networks. In section 3, we describe our proposed method to modelling information diffusion based on a nonlinear dynamical system (NLDS) that accurately models virus propagations in epidemiology. In section 4, we evaluate the accuracy of our approach by conducting experiments over real network. In section 5, we investigate the influence measurement and ranking problem based on our proposed model. Finally, conclusions and future work direction are given in section 6.

2.0 Related Works

In this section, we provide a review of recent literature on theoretical models of diffusion process in social networks. We survey and compare different approaches of measuring a node's influence and selecting influential nodes.

2.1 Diffusion Model

Models of diffusion process in social networks have been studied in various areas including epidemiology (Chakrabarti *et al.*, 2008), sociology (Granovetter, 1978), and marketing (Domingos and Richardson, 2001; Kempe, Kleinberg and Tardos, 2003). Many empirical studies on diffusion process (Leskovec, Adamic and Huberman, 2006; Backstrom *et al.*, 2006; Dasgupta *et al.*, 2008) have examined the question of how the probability of adopting new behaviour p changes as the number of friends adopting the behaviour k increases. Their results (Backstrom *et al.*, 2006) show that the plot of p versus k exhibits a similar diminishing returns effect in which the curve continues increasing, but more and more slowly, even for relatively large number of k. Building upon the empirical findings, diffusion models formulate assumptions on how individuals respond to their friends' influence and further describe the way influence flows through the network. Here we focus on the operational models that explicitly represent the step-by-step dynamics of adoption. Typically it assumes the dynamic process unfolds in discrete time unit, with each node following certain probabilistic rule (Kempe, Kleinberg and Tardos, 2003). For instance, an individual will adopt a new product or service when a certain threshold fraction of neighbours have already adopted (Granovetter, 1978). A set of nodes are chosen to be initial active set which

corresponds to the early adopters of the products or services. The active set count at the end of the process is the number of nodes reached by the initial active set.

Among the many proposed models for diffusion process, two have garnered wide acceptance. In the Linear Threshold Model (Granovetter, 1978), each node is assigned a randomly chosen threshold, representing the fraction of neighbours required for it to adopt the new behaviour. A weight is assigned on each edge, indicating the extent of the influence. A node will adopt the behaviour if sum of the weights of its neighbours that have already adopted the behaviour is greater than its threshold value. The other popular diffusion model is the Independent Cascade Model (Goldenberg, Libai and Mullen, 2001) a probabilistic model in which a node catches the behaviour from its neighbours. In this model, when a node first becomes active it gives a single chance to activate its inactive neighbours with a probability - a parameter of the system. There are many extensions of the two basic diffusion models. For instance, Kempe, Kleinberg and Tardos (2005) proposed the decreasing cascade model to incorporate the idea that a node's receptiveness to influence depends on the past history of interactions with its neighbours. In their model, a node's probability of being activated is a function of the set of neighbours have already tried and failed to influence it.

Unlike the discrete-time diffusion models discussed above, Song *et al.* (2007) proposed a continuous-time diffusion model based on diffusion rate. The diffusion rate captures how efficiently the information can diffuse among the users in the network. By leveraging the diffusion rate, their model can predict how likely the information will propagate from a specific sender to a specific receiver during a certain time period. Also it can estimate the expected time for information diffusion to reach a specific user in a network. Subsequently, they propose a DiffusionRank algorithm that ranks users based on how quickly information will flow to them.

2.2 Measuring influence and identifying influencer

Measuring influence and identify influential nodes in a network is important in many social network analysis applications. In the viral marketing context, we can target influential customers to spread viral marketing campaigns. The degree and centrality-based heuristics are commonly used in the sociology literature as estimates of a node’s influence. As mentioned earlier, there are some problems with these centrality-based heuristics. According to Kempe, Kleinberg and Tardos (2003) neither of the heuristics incorporates the fact that many of the most central nodes may be clustered, so that

targeting all of them for diffusion process is unnecessary. Link topological ranking measures such as HITS (Kleinberg, 1999) or PageRank (Brin and Page, 1998) have provided a way of measuring the authoritativeness of nodes within a network. As a variant of the eigenvector centrality measure, the intuition behind PageRank is that the importance of a web page in a network is proportional to the combined importance of its neighbours. A critical question is which of these measures is best to select marketing campaign. Kiss *et al.* (2008) compare different centrality measures for the diffusion of marketing messages. They found a significant lift when using central customers in message diffusion, but also found difference in the various centrality measures depending on the underlying network topology and diffusion process. In most cases the simple out-degree centrality outperforms almost all other measures.

3.0 Proposed Method

In this section, we describe our approach for modelling diffusion processes based on a non-linear dynamical system (NLDS) that accurately models virus propagations in epidemiology (Chakrabarti *et al.*, 2008). Leskovec, Adamic and Huberman (2006) argue that the process of new products or services diffuse through customer network is very similar to the transmission of infected diseases. Like the models we have discussed above, there is explicit notion of dynamics or time in our model. It can tell us the probability that each node is activated at some point during the process and say nothing about the particular order in which the activation occur. The calculation of the probability is based on probability theory. For instance, the probability of node v is activated at current step is the probability of the event node v is not activated at previous step and v get infected from its neighbours happens at the same time. The calculation proceeds step by step until the increment of the sum of all probabilities is less than one, which means the number of nodes reached by the initial active set has been obtained.

We start by providing some definitions for the model. Follow up to the Independent Cascade Model, we call nodes that adopt a product or service is being active and inactive otherwise. The social network is represented by a graph $G(V,E)$, where V is the set of nodes and E is the set of edges. In a customer network, nodes represent customers and edges represent the relationships between them. Let $N(v)$ denotes the set of neighbours of node v. Let A denotes the set of initial active nodes at the

beginning of the diffusion process, and it corresponds to the early adopters of a product or service. Let the probability that a node v is activated at time step t by $P_{v,t}$. Clearly, for all nodes in the initial active set $P_{v,0}$ is 1 and $P_{v,t}$ is 0 for the time step afterwards. For the rest of nodes in the network $P_{v,0}$ is equal to 0. Let θ be the infection catch rate on a link connected by an infected node. Note in contrast to a constant virus birth rate in (Chakrabarti *et al.*, 2008), in our calculation this value will be a variable depending on the past history of the process.

Let's start with the compute of node v's probability to become active at time step t. Considering node v's neighbour node w, node v has no chances of getting infected from node w is either because node w is inactive at previous time step or node w is active but failed to infect node v through the link they are connected by with probability $1-\theta$. Therefore the probability that node v has no chances of getting infected from w is $P_{w,t-1}(1-\theta)+(1-P_{w,t-1})$ which is $1-\theta * P_{w,t-1}$. It's the probability of the event that node w is active at time step $t-1$ and succeeded in infecting node v through the link they are connected by has not happened. Assuming the behaviour that each neighbour attempting to infect node v is independent of each other, hence the probability of node v has a chance of getting infected from any of its neighbours is:

$$1-\prod_{w\in N(v)}(1-\theta * P_{w,t-1}) \tag{0.1}$$

Hence node v becomes active at time step t if node v is inactive at time step $t-1$ and node v has a chance of getting infected from its neighbours:

$$P_{v,t}=(1-P_{v,t-1})*(1-\prod_{w\in N(v)}(1-\theta * P_{w,t-1})) \tag{0.2}$$

In epidemiology a susceptible individual can become infective on contact with another infective individual, and then heal herself with some probability to become susceptible again (Chakrabarti *et al.*, 2008). In this paper we only focus on the case where an active node can not be switched back to be inactive as this scenario is more common in business analytics. Taking customer churn as an example, companies typically consider churner customers that come back to the network as new customers.

In (Chakrabarti *et al.*, 2008), the virus birth rate θ is a constant representing the infectiousness of the disease. It is frequently assumed in epidemic models that individuals have equal probability of being infected every time they interact. According to Leskovec, Adamic and Huberman (2006) this may not be right. Through

observing the propagation of recommendations on a person-to-person recommendation network they found out that the probability of activation decreases with repeated interactions. For instance, if one of your friends recommended you to buy a product and you didn't buy it, the next time your friends recommended it makes sense that you are less likely to buy it. This observation is somehow consistent with the decreasing cascade model (Kempe, Kleinberg and Tardos, 2005), in which a node's propensity for being activated may change as a function of which of its neighbours have already attempted (and failed) to influence it.

Inspired by these studies, we attempt to encode the rule that the effectiveness of the influence through a particular link changes as the calculation unfolds and it depends on the past history of interactions. When a node first tries to influence its neighbour - its probability of being activated is not equal to 0 – we start to keep track of the number of trial times. Let k denotes the number of trial times. Let the probability that node v attempts to infect node w through their link by the first time be p_1, and by second and afterwards times be p_2, then

$$\theta = \begin{cases} p_1, k = 1 \\ (1-p_1)(1-p_2)^{k-2} p_{2,} \forall k \geq 2 \end{cases} \tag{0.3}$$

The reason we distinguish the first trial from the rest is motivated by a generalization of Independent Cascade Model (Kerchove *et al.*, 2009) that considering different probabilities for being infected depending on the number of contacts with the information. Their results show that first and subsequent trials play different roles in the propagation process.

Given the network structure and specified value of p_1 and p_2 we can calculate the probability for each node to get activated at every time step with the specified initial activate set. The sum of all probabilities values will keep increasing as the calculation proceeds. When the increment is less than one the calculation will terminate, as that means the expected infected nodes count has been obtained. Meanwhile, when a node's probability of getting infected is less than the value of one divided by size of the network it will be considered as negligible, which means we no longer calculate its probability in the following time steps.

4.0 Experimental Evaluations

Having described our proposed method, we will focus on understanding its behaviour in practice. Following up to the lead in (Chakrabarti *et al.*, 2008), using real network data we evaluate the accuracy of our approach of modelling information diffusion by comparing its prediction performance against diffusion simulations (Kempe, Kleinberg and Tardos, 2003; Estevez, Vera and Saito, 2007). We examine the time evolution of the infected node count at both calculation and simulation. Experiments show that our approach yields very close results to the simulations.

4.1 The network data

In our experiment, two different size data sets of scientific coauthorship network were tested. It has been argued extensively that coauthorship networks capture many of the key features of social networks more generally (Kempe, Kleinberg and Tardos, 2003). The first one is a coauthorship network of scientists working on network theory and experiment, as compiled by M. Newman in May 2006. It contains 1589 nodes and 4331 edges. This data set will be referred as NS dataset in our experiments. The second one is a weighted coauthorship network between scientists posting preprints on the High-Energy Theory E-Print Archive. There are 8361 nodes and 15751 edges in the network. It consisted of 581 connected components, and the number of nodes in the largest connected component is 5835. It's a scale-free network with a power-law degree distribution. This data set will be referred as Hep-th dataset in our experiments. Both of the networks were obtained from Mark Newman's network data collection.

4.2 The experiments

We measure the diffusion process by examining the time evolution of the infected nodes count as the dynamic process unfolds. The diffusion simulation was used as a baseline to validate the accuracy of our approach in modelling diffusion process. More specifically, we keep recode of the overall infected nodes count in the network at each time step in both cases, then check whether they are close to each other approximately. In the calculation case, the overall infected nodes count refers to the integer value of the sum of probabilities of all nodes in the network at current time step. In the simulation case, at each random process the overall infected nodes count at each time step was stored. In (Kempe, Kleinberg and Tardos, 2003), the random process will

reach equilibrium after 10k simulations. Our result is generally consistent with that. Experiments indicate that the result of 10k times is comparable to 100k times or more. The overall infected nodes count at each step then estimated by get a mean value of the 10k simulation results.

Meanwhile we make sure the experiments are conducted with the same set of initial active nodes on a given network topology. Parameters p_1 and p_2 should be specified in advance like in the Independent Cascade Model. The value of p_1 should be the same as the universal diffusion rate in Independent Cascade Model, with typically value 10%, and the value of p_2 should be smaller than p_1.

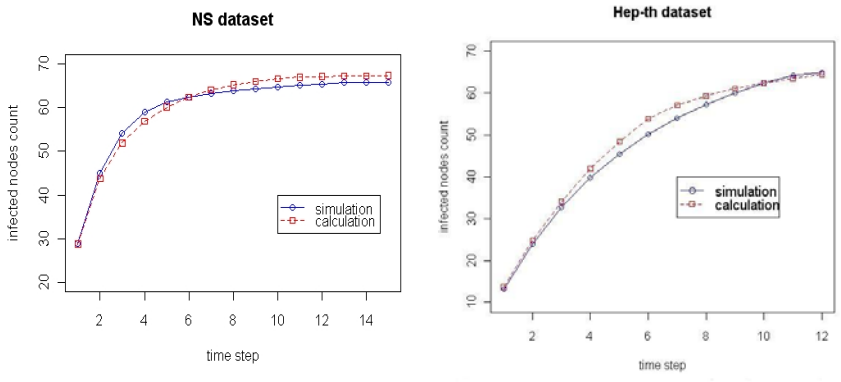

Figure 1. **The plots show the time evolution of infected nodes count in the diffusion process with both simulation and calculation. The right is the NS dataset and left is Hep-th dataset.**

The simulation was conducted using the Independent Cascade Model. In the NS dataset, the universal diffusion rate for the ICM simulations is 10%, and the p_1, p_2 value for our proposed estimation method is 10% and 7.8%, respectively. In the Hep-th dataset, the universal diffusion rate for the ICM simulations is 9%, and the p_1, p_2 value for our proposed estimation method is 9% and 8%, respectively. The initial active set A was chosen randomly. With different initial active sets, the results were almost the same – two curves are similar two each other. In both datasets the two curves exhibit qualitatively similar shapes, dominated by a diminishing returns property in which the curve continues increasing, but more and more slowly and eventually flattened. As shown, our method nicely tracks the simulation results.

5.0 Influence Measurements and Ranking

In this section, we look into the possibility of measuring nodes' influence value and identify influential nodes in social networks based on our proposed method. As

mentioned earlier, the number of nodes reached by an initial target set in the diffusion process should be an indicator of the importance or influence of the initial target set. Last section we demonstrate the accuracy of method in modelling diffusion process. Hence we can use the sum of all probability values as an approximate estimation to the number of nodes reached in the diffusion process. The algorithm we used to calculate this quantity can be described as follows: The inputs needed are the network structure, proper parameters setting and initial target set. Starting with time step 1, calculate each node in the network's probability to get activated at current time step and sum up all the probability values, and then go to next time step. Repeat this process until the increment of all probability values is less than one. Here we calculate this quantity (For simplicity, we refer to it as influence value) for each node in the NS dataset, and further compare the influence value calculated with node degree.

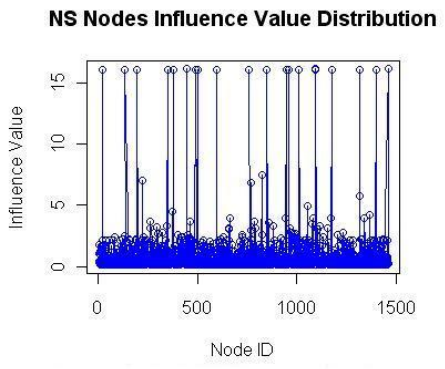

Figure 2. The plot shows the distribution of influence value for all nodes in the NS dataset.

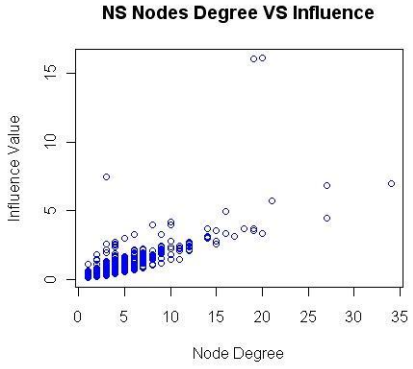

Figure 3. The plot compares the influence value and node degree for all nodes in the NS dataset.

As shown in Figure 2, in the NS dataset most of the nodes can only influence one or two nodes in the diffusion process, while only a few nodes can influence more than two nodes. Intuitively, it seems beneficial to target those nodes that can influence more than two nodes in the diffusion process to spread viral marketing campaigns. As shown in Figure 3, node degree and influence value calculated are not fully correlated. For instance, there is one node with degree 4 and influence value around 8. High degree does not necessarily mean high influence value, and vice versa. This implies that the influence value metrics provides ranking methods that in general extract nontrivial nodes as influential nodes.

6.0 Conclusion

In this paper, we extend an epidemic model that accurately models virus propagation. With proper parameters setting on a given network topology, our method can numerically calculate each node's probability to get infected when a set of nodes has been initially activated. By comparing its predicting performance with diffusion simulations, we validate the accuracy of our proposed method in modelling diffusion process. When it comes to compute the number of nodes reached by set of initial active nodes, our model can give a suitable estimation to this quantity. Using the number of nodes reached in the diffusion process as an influence measure, experiments results show that our proposed method provides ways of extracting nontrivial nodes as influential nodes.

The development of theoretical models for diffusion process still remains to be an open question. First of all, are these models discussed above correctly captured the way influence spreads through real network? All the models take a snapshot of the network, and then operate upon this fixed snapshot. No dynamic aspects or network evolution involves – it does not consider the network growths. Also all the models unfold in discrete time step with each node following certain probabilistic rule, and it uses this rule to incorporate information from its neighbour over time. Whereas the dependence of probability of adopting behaviours on number of friends adopted expressed in this way reflects an aggregate property of the full population, and does not imply anything about any particular individual's respond to their friends' behaviours (Kleinberg, 2007). Secondly, the way such dynamic process is affected by the network structure is still poorly understood (Kiss and Bichler, 2008). How adoption probability depends on the structural properties of a node's network neighbours? What role does weak and strong ties play in the dynamic process? Is information propagates more quickly on a dense network?

While the theoretical models address the question of how influence spreads in a network, they are based on assumed rather than measured influence effects (Leskovec, Adamic and Huberman, 2006). According to Backstrom (2006) it has to date been easier to explore such models theoretically than to obtain reasonable estimates for them empirically on large-scale data. Our future work directions include obtaining actual information diffusion data and observing how influence propagates in real

network. Therefore we could develop ways to infer or estimate relevant model parameters with the historical diffusion data (Saito, Nakano and Kimura, 2008). With the support of the empirical findings we could make more general assumptions on how individuals respond to friends' influence, which leads to a closer integration of the theoretical models to the empirical results.

References

Brin, S. and Page, L. (2008) The Anatomy of a Large-Scale Hypertextual Web Search Engine, Computer Networks and ISDN Systems, Vol. 30, No. 1-7, pp. 107-117.

Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. (2006) Group formation in large social networks: membership, growth, and evolution, In KDD '06: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pages 44-54, New York, NY, USA.

Chakrabarti, D., Wang, Y., Wang, C., Leskovec, J. and Faloutsos, C. (2008) Epidemic Thresholds in Real Networks, ACM Transactions on Information and System Security, Vol. 10, No. 4.

Domingos, P. and Richardson, M. (2001) Mining the network value of customers, In KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pages 57-66, New York, NY, USA.

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A. and Joshi, A. (2008) Social Ties and their Relevance to Churn in Mobile Telecom Networks, In: Proc. 11th International Conference on Extending Database Technology, pp. 668-677.

Doyle, S. (2008) Social network analysis in the Telco Sector- marketing applications, Journal of Database Marketing & Customer Strategy Management, Vol. 15, No. 2, pp. 130-134.

de Kerchove, C., Krings, G., Lambiotte, R. and van Dooren, P. (2009) The Role of Second Trials in Cascades of Information over Networks, Physical Review E, Vol. 79, No. 1.

Estevez, P.A., Vera, P., Saito, K. (2007) Selecting the Most Influential Nodes in Social Networks, In: International Joint Conference on Neural Networks, pp. 2397-2402

Granovetter, M. (1978) Threshold models of collective behavior, The American Journal of Sociology, 83(6):1420-1443.

Gabbott, M. and Hogg, G. (1994) Consumer behaviour and services: a review, Journal of Marketing Management, 10(4):311-324.

Goldenberg, J., Libai, B., and Mullen, E. (2001) Talk of the network: a complex systems look at the underlying process of word-of-mouth, Marketing Letters, 12(3):211-223.

Kleinberg, J. (1999) Authoritative sources in a hyperlinked environment, Journal of the ACM 46 (5): 604–632.

Kempe, D., Kleinberg, J., and Tardos, Ã. (2003) Maximizing the spread of influence through a social network, In KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 137-146, New York, NY, USA.

Kempe, D., Kleinberg, J., and Tardos, Ã. (2005) Influential nodes in a diffusion model for social networks, In: Proc. 32nd ICALP, pages 1127-1138.
Kleinberg, J. (2007) *Cascading Behavior in Networks: Algorithmic and Economic Issues,* Cambridge University Press.
Kiss, C. and Bichler, M. (2008) Identification of Influencers - Measuring Influence in Customer Networks, Decision Support Systems, Vol. 46, No. 1, pp. 233-253.
Leskovec, J., Adamic, L. A., and Huberman, B. A. (2006) The dynamics of viral marketing, In EC '06: Proceedings of the 7th ACM conference on Electronic commerce, pages 228-237, New York, NY, USA.
Song, X., Chi, Y., Hino, K., and Tseng, B. L. (2007) Information flow modeling based on diffusion rate for prediction and ranking, In WWW '07: Proceedings of the 16th international conference on World Wide Web, pages 191-200, New York, NY, USA.
Saito, K., Nakano, R. and Kimura, M. (2008) Prediction of Information Diffusion Probabilities for Independent Cascade Model, KES (3), pp. 67-75.