A Unified Model of Influence in Social Networks?

Ajitesh Srivastava
Department of Computer Science
University of Southern California
Los Angeles, USA
ajiteshs@usc.edu

Charalampos Chelmis, Viktor K. Prasanna
Ming Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, USA
{*chelmis, prasanna*}*@usc.edu*

***Abstract*—It is difficult to obtain accurate prediction results of cascades over online social networks, therefore a variety of diffusion models have been proposed in the literature to simulate diffusion processes instead. We argue that such models require extensive simulation results to produce good estimates of future spreads, while at the same time requiring training over observed data to learn the parameters that they incorporate into the various influence mechanisms that drive diffusion. In this work, we take a complimentary approach. We present a generalized, analytical model of influence in social networks that captures social influence at various levels of granularity, ranging from pairwise influence, to local neighborhood, to the general population, and external events, therefore capturing the complex dynamics of human behavior. Commonly used diffusion models in social networks can be reduced to special cases of our model, by carefully defining their parameters. Our goal is to provide a closed-form expression for the probability of infection for every node in an arbitrary, directed network at any time t. However prior work in the literature has shown that exact computation of infection probabilities is #P-hard. We make an independence assumption about the infection events of a node's incoming neighbors, which results our formula being an approximation. We quantitatively evaluate the approximation quality of our analytical solution as compared to numerous popular diffusion models on a real-world dataset and a series of synthetic graphs.**

***Keywords*-analytical framework; computational models; diffusion models; dynamics; evolutionary models; influence**

I. Introduction

Influence analytics and diffusion prediction in online social networks have been important for many domains from marketing to public health. With the tremendous increase in the volume of data, network sizes reach millions of nodes, restricting the applicability of computational models for diffusion prediction. Prior work on diffusion processes [1], [2], [3] includes studying the diffusion of innovations [4], [5] and word-of-mouth recommendations [6]. Existing models of spreading processes in networks attempt to model diffusion as a result of social influence, i.e., the more influential a user is the wider the spread [7]. Diffusion is modeled using a network structure with static or dynamic edge probabilities [6], [8], which are estimated from past observational data [1]. According to such models, each node independently infects its neighbors with some probability, and each infected node then propagates the infection in the network. Even though this process captures individual influence (i.e., node-to-node), it ignores social influence effects which appear as a result of neighborhood or global pressure [5]. [7] proposed to mediate this problem by incorporating the notion of social capital to characterize the network effect in the influence process, whereas [5] presented agent-based computational models, which quantified pairwise influence and global dynamics in the spread of technology adoption at the workplace. Two of the most widely used diffusion models are the Linear Threshold Model (LTM) [9], and the Independent Cascade Model (ICM) [6].

Typically proposed methodologies for influence calculation and models of diffusion need extensive simulation results to be evaluated, usually by means of statistical analysis. In fact, it has been shown that exact computation of infection probabilities is #P-hard [10]. Instead, we devise a novel formulation of progressive diffusion with minimum computational complexity. We provide a generalized, analytical solution to the diffusion mechanism that comprises of two processes unfolding over the network simultaneously: (a) pairwise influence, and (b) pressure from collective dynamics. Particularly, our work introduces an important dimension to the diffusion process, which in our case explicitly encompasses not only pairwise influence, but also local neighborhood effects, aggregate social behavior, and external factors, or a combination of the above.

Our methodology is vertex-centric, i.e., models each user separately, offering great flexibility in terms of modeling personalized influence functions, and allows for the use of time-dependent influence functions. Note, that in this work, we are not concerned with learning the parameters that drive the spread of infection from observational data. While this aspect is important, it is outside of the scope of this paper.

To the best of our knowledge, our work is the first to enable analytical computation of complex, non-linear phenomena like influence, while considering multiple factors that can change over time, without requiring extensive simulation runs to estimate the propagation probabilities at the steady state. Our formula explicitly and formally unites a rich class of popular diffusion processes in social networks [9], [6], [5] as special cases.

II. Analytical Model of Influence

A. Unified Model of Influence

We model a social network as a directed graph $G = (V, E)$, where a node $v \in V$ represents an individual, and edge $(v, u) \in E$ exists if v interacts with u (in our context v influences u). For every node v, we define the set of incoming neighbors $N_i(v) = \{u|(u, v) \in E\}$, and the set of outgoing neighbors $N_o(v) = \{u|(v, u) \in E\}$. Our goal is to model the probability of infection for every node in the network at any time t. We start with a seed set $S \subset V$ of infected nodes at time $t = 0$. The infection process proceeds in discrete time steps, in which two types of influence unfold over the network [5]. First, each infected node v attempts to infect its neighbors (*individual* influence). Each attempt of infecting node $u \in N_o(v)$ has a chance of success, but the probability of infection $p_{(v,u)}(t)$ is pairwise and may change over time. We assume infection attempts from different neighbors to be independent. Second, each susceptible node u can be infected with probability $r_u(t)$, independent of individual influence. Such *collective* influence may include external factors [11], or external sources of exposure [13], or the status of the incoming neighborhood of u [5].

We note that: (i) function $r_u(t)$ is node specific, and may be time dependent, (ii) there may be arbitrary number of collective influence attempts on a node, as we assume $r_u(t)$ is not conditioned upon the node already having undergone a collective influence attempt or not. The process repeats until a pre-specified stopping criterion is satisfied (e.g., number of time steps elapsed, or fraction of infected nodes has exceeded some number).

B. Infection Probability Formula Under the Unified Model

Let $B_{u,t}$ represent the probability of infection of node u by the time t. Initial values $\{B_{u,0}\}$ are either 0 or 1 depending on the membership of u in the seed set. Let $E_{v,t}$ denote the indicator variable, which is 1 if node v is infected by the time t, 0 otherwise. To find the probability of a node u being infected at time t, we consider an arbitrary ordering of its incoming neighbor set $N_i(u)$: $< v_1, v_2, \ldots, v_n >$. Based on this, we define *zero state probability* at time $t - 1$: $P^0_{s_n,s_{n-1},\ldots,s_1}$, where superscript 0 denotes $E_{u,t-1} = 0$. The subscript is a vector, which elements s_i denote the value of $E_{v_i,t-1}$, and can take values in $\{0, 1, *\}$. $s_i = 0$ represents $E_{v_i,t-1} = 0$, $s_i = 1$ denotes $E_{v_i,t-1} = 1$, and $s_i = *$ indicates marginalization over the state of v_i, i.e., '$E_{v_i,t-1} = 0$ or 1'. For instance, for a node u with four neighbors, $P^0_{0,1,*,1}$ denotes the probability $P(E_{u,t-1} = 0, E_{v_4,t-1} = 0, E_{v_3,t-1} = 1, E_{v_1,t-1} = 1)$. We begin by calculating $B_{u,t}$ in the special case of G being a tree, i.e., each node has at most one incoming neighbor.

Corollary 1: The infection probability of node u with parent v in a tree is given by:

$$\begin{aligned} B_{u,t} = 1 - (1 - r_u(t))\big((1 - p_{v,u}(t))(1 - B_{u,t-1}) \\ + p_{v,u}(t)(1 - B_{v,t-1}) \prod_{k=1}^{t-1}(1 - r_u(k))\big) \,. \end{aligned} \quad (1)$$

Proof: The probability of node u not being infected by time t is $P(E_{u,t} = 0) = 1 - B_{u,t}$. Either one of two things must have happened for u not to be infected by time t. First, state $E_{u,t} = 0$ was reached from state $(E_{v,t-1} = 0, E_{u,t-1} = 0)$ if and only if collective influence $r_u(t)$ failed to infect u at time t. Intuitively, when the parent of u was not infected at time$t - 1$, the only chance for u to be infected at time t is through collective influence $r_u(t)$, with probability $1 - r_u(t)$. Second, state $E_{u,t} = 0$ was reached from state $(E_{v,t-1} = 1, E_{u,t-1} = 0)$, i.e., when the parent of u was infected at time $t - 1$, if and only if collective influence was unsuccessful, and furthermore v failed to infect u. As the two processes are independent, this can happen with probability $(1 - r_u(t))(1 - p_{v,u}(t))$. It follows that,

$$\begin{aligned} 1 - B_{u,t} &= P^0_1(1 - r_u(t))(1 - p_{v,u}(t)) + P^0_0(1 - r_u(t)) \\ &= (1 - r_u(t))(P^0_1(1 - p_{v,u}(t)) + P^0_0) \\ &= (1 - r_u(t))((P^0_* - P^0_0)(1 - p_{v,u}(t)) + P^0_0) \\ &= (1 - r_u(t))(P^0_*(1 - p_{v,u}(t)) + P^0_0 p_{v,u}(t)) \,, \end{aligned} \quad (2)$$

where $P^0_* = P(E_{u,t-1} = 0) = 1 - B_{u,t-1}$ and $P^0_0 = P(E_{u,t-1} = 0, E_{v,t-1} = 0)$. This means v and u were both susceptible at time $t - 1$. If v was also not infected by the time $t - 1$, u can only be susceptible because all collective influence till that time failed, i.e., $P^0_0 = (1 - B_{v,t-1}) \prod_{k=0}^{t-1}(1 - r_u(k))$. We set $r_u(0) = 1$ if $u \in S$, 0 otherwise. Substituting the values of P^0_* and P^0_0 in Equation 2, results in Equation 1. ■

Next, we extend Equation 1 to a graph of any type. Without loss of generality, we focus on directed graphs, as undirected graphs can be converted into their directed equivalent.

Lemma 1: The probability of a node u not being infected by the time t is related to the zero state probabilities as follows

$$\begin{aligned} &1 - B_{u,t} = (1 - r_u(t)) \\ &\sum_{s_i \in \{0,*\}} \left(P^0_{s_n,s_{n-1},\ldots,s_1} \prod_{i=1}^{n} (1 - p_{v_i,u}(t))^{\delta_{s_i,*}} \prod_{i=1}^{n} p_{v_i,u}(t)^{\delta_{s_i,0}} \right) . \end{aligned} \quad (3)$$

where $\delta_{a,b} = 1$ only if $a = b$, 0 otherwise, is the Kronecker delta function.

Proof: When the number of incoming neighbors is one, Lemma 1 follows from Equation 2. Now, suppose the statement is true for $k \geq 1$ parents. Consider a sequence $\mathbf{x_k} =< s_k, s_{k-1}, \ldots, s_1 >$. We look at the new terms that are added due to the inclusion of v_{k+1}. For ease of notation, let $D(\mathbf{x_k}) = (1 - r_u(t)) \prod_{i=1}^{k}(1 - p_{v_i,u}(t))^{\delta_{s_i,*}} \prod_{i=1}^{k} p_{v_i,u}(t)^{\delta_{s_i,0}}$. Equation 3 can be rewritten as

$$1 - B_{u,t} = \sum_{\mathbf{x_n}} P^0_{\mathbf{x_n}} D(\mathbf{x_n})\,. \tag{4}$$

We have assumed that this is true for $n = k$. The addition of v_{k+1} affects $P(E_{u,t})$ in ways similar to those discussed in Corollary 1, i.e., if $E_{v_{k+1},t-1} = 1$, then this new node fails to infect u with probability $(1 - p_{v_{k+1},u}(t))$. On the other hand, if $E_{v_{k+1},t-1} = 0$, node v_{k+1} does not have the ability to infect. Formally, the new terms added are:

$$\begin{aligned}
&P^0_{1,\mathbf{x_k}}(1 - p_{v_{k+1},u}(t))D(\mathbf{x_k}) + P^0_{0,\mathbf{x_k}}D(\mathbf{x_k}) \\
=&(P^0_{*,\mathbf{x_k}} - P^0_{0,\mathbf{x_k}})(1 - p_{v_{k+1},u}(t))D(\mathbf{x_k}) + P^0_{0,\mathbf{x_k}}D(\mathbf{x_k}) \\
=&P^0_{*,\mathbf{x_k}}(1 - p_{v_{k+1},u}(t))D(\mathbf{x_k}) + P^0_{0,\mathbf{x_k}}p_{v_{k+1},u}(t)D(\mathbf{x_k}) \\
=&P^0_{*,\mathbf{x_k}}D(*,\mathbf{x_k}) + P^0_{0,\mathbf{x_k}}D(0,\mathbf{x_k})\,,
\end{aligned}$$

which would generate the required terms in the right hand side of Equation 4, when $n = k+1$. This indicates that the statement is true for $k+1$ incoming neighbors. By induction, Lemma 1 is true $\forall n$. ■

Theorem 1: An approximate probability of infection is given by the recurrence relation:

$$\begin{aligned}
B_{u,t} = 1 - \Bigg[&(1 - B_{u,t-1})\Bigg(\prod_{v \in N_i(u)}(1 - p_{v,u}(t)B_{v,t-1})\Bigg) \\
&+ \Bigg(\prod_{v \in N_i(u)} p_{v,u}(t)(1 - B_{v,t-1})\Bigg) \\
&\Bigg(\prod_{k=1}^{t-1}(1 - r_u(k)) - 1 + B_{u,t-1}\Bigg)\Bigg](1 - r_u(t))\,.
\end{aligned} \tag{5}$$

The above formula is an approximation due to the assumption that the incoming neighbors of a given node u are independently influenced, i.e., for two incoming neighbors v_i and v_j, events $E_{v_i,t-1} = 0$ and $E_{v_j,t-1} = 0$ are independent. This provides us with a closed-form expression. Next, we proceed with proving the Theorem.

Proof: We attempt to find the zero state probabilities for sequence $\mathbf{x_n}$. When $\mathbf{x_n} =< 0, 0, \ldots, 0 >$, u and all nodes in $N_i(u)$ are susceptible, which means that collective influence till $t-1$ was unsuccessful. Further, at $k = 0$, $r_u(t) = 1$ for $u \in S$. In this case,

$$P^0_{0,0,\ldots,0} = \prod_{k=0}^{t-1}(1 - r_u(k)) \prod_{v \in N_i(u)}(1 - B_{v,t-1})\,. \tag{6}$$

Any other sequence $\mathbf{x_n}$, which consists of at least one $*$ in the i-th position, represents the state of u being not infected by the state of its i-th neighbor. Given the state of u's neighbors, the conditional probability of u not being infected is $1 - B_{u,t-1}$. The zero state probability is then computed as follows

$$P^0_{\mathbf{x_n}} = (1 - B_{u,t-1}) \prod_{s_i=0}(1 - B_{v_i,t-1})\,. \tag{7}$$

Combining Equations 4, 6 and 7, results in the following:

$$\begin{aligned}
&\frac{1 - B_{u,t}}{1 - r_u(t)} = (1 - B_{u,t-1})\Bigg(\prod_{v_j \in N_i(u)}(1 - p_{v_j,u}(t))\Bigg) \\
&\Bigg(\sum_{\mathbf{x_n}}\prod_{s_j=0}\frac{(1 - B_{v_j,t-1})p_{v_j,u}(t)}{1 - p_{v_j,u}(t)}\Bigg) + \\
&\Bigg(B_{u,t-1} - 1 + \prod_{k}(1 - r_u(t))\Bigg)\Bigg(\prod_{v_j}(1 - B_{v_j,t-1})p_{v_j,u}(t)\Bigg)\,.
\end{aligned}$$

After simplification, the above equation reduces to Equation 5. This step completes the proof. ■

C. *Complexity Analysis*

The recurrence relation in Equation 5 requires inspection of all incoming links to a node u, $|N_i(u)|$, at every time step. Therefore, in order to evaluate Equation 5 for all nodes for t time steps, the number of operations required is $t\sum_u O(|N_i(u)|) = O(|E|t)$.

D. *Reduction to Other Models*

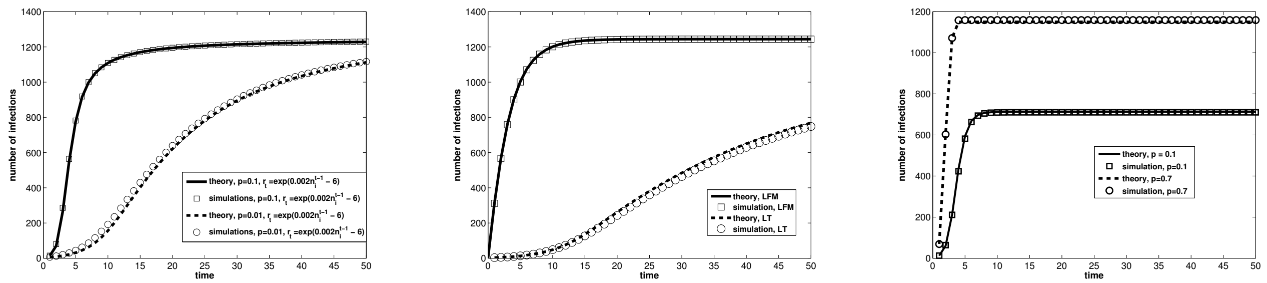
Our analytical formula of influence in social networks, offers great flexibility in terms of modeling a variety of diffusion processes. It can be shown that popular diffusion models [4], [6], [8], [1], [5] can be reduced to special cases of the Unified Model, by carefully defining the individual influence probabilities and collective influence functions.

III. Experiments

We illustrate the ability of our Unified Model to capture real-life behavior on a real-world dataset $(1,244$ users and $28,343$ directed links) from Digg [14]. We compare the predicted values obtained by Theorem 1 to $1,000$ simulations of popular diffusion models on the task of information diffusion. For each model, we start with a seed set of two infected nodes. Table I summarizes the set of parameters used in our experiments. Figure 1 shows the results. Our findings imply that Equation 5 is able to accurately predict the expected epidemics forecasted by the rest of the models without extensive numerical simulations.

IV. Conclusion

In this work, we have proposed a novel, general analytical framework for influence calculation in social networks, which does not require extensive simulations. In this framework, each node has its own individual function of collective influence and pairwise influence functions for each neighboring node. Both functions vary with time, thus



(a) Complex Contagion Model (b) Threshold Models (c) Independent Cascade Model

Figure 1. Agreement of simulation and theory for the three models for Digg1k dataset.

Table I
PARAMETERS USED IN THE EXPERIMENTAL VALIDATION ON DIGG FOLLOWER GRAPH

parameter set 1	CCM	$p = 0.1, r(t) = exp(0.002n_i^{t-1} - 6)$
	GLT	$f(In(u,t), \vec{b_u}) = \sum_{v \in In(u,t)} b_{v,u}$
	ICM	$p = 0.1$
parameter set 2	CCM	$p = 0.01, r(t) = exp(0.002n_i^{t-1} - 6)$
	GLT	$f(In(u,t), \vec{b_u}) = \frac{\exp(\lvert In(u,t,\vec{b_u})\rvert)}{1+\exp(\lvert In(u,t,\vec{b_u})\rvert)}$
	ICM	$p = 0.7$

making our framework directly applicable to a plethora of situations. We have validated our results on a real-world social network.

ACKNOWLEDGMENT

This work is supported by Chevron Corp. under the joint project, Center for Interactive Smart Oilfield Technologies (CiSoft), at the University of Southern California.

REFERENCES

[1] C. Budak, D. Agrawal, and A. El Abbadi, "Diffusion of information in social networks: Is it all local?" in *2012 IEEE 12th International Conference on Data Mining (ICDM)*, 2012, pp. 121–130.

[2] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *Proceedings of the 21st international conference on World Wide Web*. New York, NY, USA: ACM, 2012, pp. 519–528.

[3] A. Hajibagheri, A. Hamzeh, and G. Sukthankar, "Modeling information diffusion and community membership using stochastic optimization," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '13. New York, NY, USA: ACM, 2013, pp. 175–182.

[4] T. W. Valente, "Social network thresholds in the diffusion of innovations," *Social Networks*, vol. 18, no. 1, pp. 69–89, 1996.

[5] C. Chelmis and V. K. Prasanna, "The role of organization hierarchy in technology adoption at the workplace," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '13. New York, NY, USA: ACM, 2013, pp. 8–15.

[6] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2003, pp. 137–146.

[7] K. Subbian, D. Sharma, Z. Wen, and J. Srivastava, "Finding influencers in networks using social capital," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '13. New York, NY, USA: ACM, 2013, pp. 592–599.

[8] J. Kleinberg, *Cascading Behavior in Networks: Algorithmic and Economic Issues*. Cambridge University Press, 2007.

[9] M. Granovetter, "Threshold models of collective behavior," *American Journal of Sociology*, vol. 83, no. 6, pp. 1420–1443, 1978.

[10] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 1029–1038.

[11] E. Abrahamson and L. Rosenkopf, "Social network effects on the extent of innovation diffusion: A computer simulation," *Organization Science*, vol. 8, no. 3, pp. 289–309, 1997.

[12] S. A. Myers, C. Zhu, and J. Leskovec, "Information diffusion and external influence in networks," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 33–41.

[13] Y.-R. Lin, J. Sun, P. Castro, R. Konuru, H. Sundaram, and A. Kelliher, "Metafac: Community discovery via relational hypergraph factorization," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09. New York, NY, USA: ACM, 2009, pp. 527–536.