Modeling the Diffusion of Preferences on Social Networks

Jing-Kai Lou*[2], Fu-Min Wang†[1], Chin-Hua Tsai‡[1], San-Chuan Hung§[1], Perng-Hwa Kung¶[1], and Shou-De Lin‖[1]

[1]Department of Computer Science and Information Engineering, National Taiwan University
[2]Department of Electrical Engineering, National Taiwan University

Abstract

The information diffusion on social networks has been studied for decades. To simplify the diffusion on social networks, most models consider the propagated information or media as single values. Representing media as single values however would not suitable for certain concernful situations such as the voter preference toward the candidates in an election. In such case, the representation would better be lists instead of single values so people can try to alter others' preference through social inference. This paper studies the diffusion of preference on social networks, which is a novel problem to solve in this direction. First, we propose a preference propagation model that can handle the diffusion of vector-type information instead of only binary or numerical values. Furthermore, we theoretically prove the convergence of diffusion with the proposed model, and that a consensus among strongly connected nodes can eventually be reached with certain conditions. We further extract relevant information from a publicly available bibliography datasets to evaluate the proposed models, while such data can further serve as a benchmark for evaluating future models of the same purpose. Lastly, we exploit the extracted data to demonstrate the usefulness of our model and compare it with other well-known diffusion strategies such as independent cascade, linear threshold, and diffusion rank. We find that our model consistently outperforms other models.

1 Introduction

With the success of viral marketing, people see the power of crowd opinions, and believe that individual options or preferences could be highly affected by acquaintances even though individuals generally possess intrinsic preferences. For instance, in an election, people would argue and even attempt to convince others for their favorite candidates. With the rise of social networking service (e.g. Facebook or Twitter) in Web 2.0 era, people would create, or reply posts to promote the positions of their favorite candidates on others' mind. In such case, it is the preference toward a set of candidates that is propagated in a social network. Up to date we have not yet seen too many computational approaches with systematic and quantifiable studies on this issue. Nevertheless, being able to model the human preference does possess its own value in the real world applications. Social scientists might wonder to what extent the opinions exchange among friends can affect each other's viewpoints toward an object. Campaign companies might inquire how to promote a candidate given a limited budget through a social network. Such questions are not easy to answer via a real-world user study, in particular when the network becomes huge.

Although the issues about information propagation on social networks have been studied for decades, many proposed models such as the independent cascade model, linear threshold model, SIR/SIS model, and heat diffusion model, unfortunately, assume the sources for propagation is either a binary value or a real number. They cannot be applied directly to solve our problem where it is a preference list that needs to be propagated on the network. The goal of our study, therefore, is to design a suitable framework that allows us to model the preference propagation on social networks.

To handle the information propagation such as the situation in election, we have realized several properties that a suitable preference propagation model require, namely hyper dimensional media, input dependent, deterministic convergence, and consensus. The properties are intuitively inspired by the natural real-world phenomena, and are summarized as the follows. First, we prefer the media (which represents preference toward different candidates) propagated throughout the process being a real valued vector that sums to one, because in the real world scenario usually each node (or individual) has equal right in casting votes. Second, the preference distribution should be affected significantly by the ini-

*kaeaura@gmail.com
†r98723077@ntu.edu.tw
‡zhichin@gmail.com
§c2016.tw@gmail.com
¶answerseeker95@gmail.com
‖sdlin@csie.ntu.edu.tw

tial intrinsic preference as well as the social network topology. Finally, we hope the propagation should converge eventually, and the common trends in real world would finally appear after a great number of interactions [11]. We will show that our model is the only one that can satisfy all properties among the existing models.

The novelty and contributions of this paper can be viewed from several different angles.

1. We design a novel information preference propagation model which focuses on the propagation of a vector instead of a single value. This model is not only simple and intuitive, but also capable of producing several meaningful real-world behavior such propagation.

2. We propose the importance properties to follow for a preference propagation model. We also assess the quality of the proposed model by proving its convergence and several other important properties.

3. We propose a novel way to obtain relevant information and ground truth from publicly available datasets to evaluate the preference propagation models. Such data can further serve as a benchmark for future models of the same purpose. Using the ground truth obtained, we then conduct experiments to demonstrate the validity of our model in predicting the change of citation preference among authors through collaboration networks.

2 Related Work

Linear Threshold Model (LT Model) [6, 8] and Independent Cascade Model (IC Model) [8, 2] are well-known cascading models, and are the foundation for a number of more sophisticated models. In the LT model, a real-value weight is propagated through the network. In the IC model, by contrast, only binary signal are considered. Kempe et al.(2003) [8] generalized the IC model by introducing a General Cascade Model. D.Grul et al.(2004) [7] and Leskovec et al.(2006) [10] proposed generative model to simulate blog essay generation based on the IC Model. These models assume nodes can turn from inactive to active given a certain probability for cascading. Based on the LT Model and the IC Model, Saito et al.(2010) [15] proposed Asynchronous Linear Threshold Model and Asynchronous Independent Cascade Model.

However, the above models are formulated on the assumption that model states are binary (active or inactive) and there is a binary signal and real-number (weight) propagating in the network. It is very different from our model which assumes the propagation of an ordered preference list.

Another influential line of research, following the success of the PageRank algorithm, is to put the propagation process in an explicit recursive mathematical form. The Heat Diffusion model [13, 18] is an approach to simulate the diffusion process. Heat Diffusion is a physics phenomenon describing heat flows from high temperature positions to low temperature positions. Using Heat Diffusion, Ma et al.(2008) [13] proposed a model to analyze candidates selection strategies for market promotion. Below is a description of diffusion. The process is formulated as:

$$\frac{f_i(t+\Delta t) - f_i(t)}{\Delta t} = \alpha \sum_{j:(v_j,v_i)\in E} (f_j(t) - f_i(t)),$$

where $f_i(t)$ is the heat of node i at time t, and α is the thermal conductivity-the heat diffusion coefficient.

In the Heat Diffusion process, each vertex receives heat from its neighbors, which is similar to the propagation phase of our model. The major difference, will be discussed in the following section, is that such model lacks a normalization phase (since it considers only the propagation of one value) and a fusion phase (because the heat itself can disappear after diffusion, so there is no need to fuse on heat diffusion model).

Inspired by these previous works, our model seeks to take the strong points of each of these approaches, namely, their focus on mimicking commonly observed social interaction characteristics such as forming consensus as well as their incorporation of structural information into the propagation process, and blend them into a more coherent framework that could be used to answer real world social problems of interest.

In 1992, Bartholdi, Tovey, and Trick [1] first studied the complexity of the process to determine needed actions by organizer to add or remove candidates to manipulate election results (where it is recognized as the classical social choice theory). However, they did not propose any model for the interactions between voters. Gibbard [5] and Satterhwaite [16] showed that every election scheme with at least three possible outcomes is subject to individual manipulation. This means the minority has a chance to manipulate the group decision to secure a preferred outcome. Gibbard and Satterhwaite also addressed the computational difficulty in minority manipulation. However, their model assumes independence of voters, which means no voter's preference can be affected by others. Nevertheless, existing studies in this direction still focus on the complexity and feasibility issues, which is very different from the goal of this paper.

Liu(2009) [12] attempted to check whether the preference distribution changes if the number of political experts in a communication network increases. They use an agent-based model for simulation. In this model, the agent maintains a binary value toward a candidate (instead of a real value or ranking) and simply propagate such values to the other agents in the nearby 3 by 3 matrix. Yoo et al. (2009) [19] proposed semi-supervised importance propagation model. Their idea is to some

Table 1: Notations

Notation	Description
V	individuals
C	candidates
n	number of individuals
k	number of candidates
p_v	preference profile vector of individual $v \in V$
s_v	preference score vector for individual $v \in V$
S	preference scoring matrix with size $\|V\| \times \|C\|$
G	social network layer

extent similar to our "fusion phase" by adding the original score into the accumulated score obtained from the neighbor. The difference between their model and ours is that theirs deal with a single value instead of a vector, and therefore do not perform the normalization over candidate scores like we do.

sectionPreference Propagation Model n individual's preference could change gradually when communicating with his or her acquaintances. To understand how such communication can affect the individual decisions, we propose a propagation model for preference in social networks. The notations used in the residual article are listed in Table 1.

2.1 Preference Propagation Model We first define a preference profile p_v of an individual v, which is a k-dimensional vector that represents v's preference toward k different candidates. The jth element in p_v is an integer in $[1, k]$ indicating this individual's preference for candidate j (smaller numbers denote higher ranks). To facilitate the operation of the preference profiles, we translate p_v into a score vector s_v for all v using the following equation:

$$s_v[i] = (k - p_v[i] + 1)/T, \forall i \in 1, 2, \ldots, k$$

where $T = k(k+1)/2$. This transformation can be regarded as a normalization process as in s_v not only the preferred candidate receives higher score but also the sum of all element equals to 1.

Using the score vector of each individual, we can create an n by k matrix $S = (s_{v_1}, s_{v_1}, \ldots, s_{v_n})^t$ denoted as the preference matrix. Next, we would like to model how the preference matrix of a given time stamp t, $S(t)$, changes after the propagation process starts. We assume the edge directions in a network G reveal the direction of influence.

The information propagates one iteration after another in our model, and each iteration consists of three phases: propagation, normalization, and fusion. In the propagation phase, each node v synchronically propagate the preference score vector s_v to the neighboring nodes. To describe such operation mathematically, we define an $n \times n$ forward transition matrix F such that the multiplication of F and $S(t)$ represents the score of each node obtained from all neighbors after this phase. We denoted it as $S_p(t)$. In details, $F = (KA)^t$, where K is a diagonal matrix with the inverse of degree of each node in the diagonal, and A is the adjacency matrix of G. Note that F is identical to the forwarding matrix of a random walk algorithm. The only difference is that F in a random walk algorithm is multiplied by a vector instead of a matrix S.

In S_p, each row represents the neighbors' accumulated preference scores toward each candidate. However, unlike S, the elements in each row of S_p do not add up to one. To ensure every individual has equal influence while casting votes, we normalize each row of S_p so its elements add up to one. Therefore, in the second phase, S_p is multiplied by a $n \times n$ diagonal normalization matrix N, where each element in the diagonal of N is equal to the sum of all elements in the corresponding row of S_p. After the second phase, we will obtain a new scoring matrix $S_n(t) = NFS(t)$.

The major difference between our propagation model and the diffusion models for electricity/heat (see Section 2 for more detail) lies in the intrinsic difference of the media that are propagated. Electricity or heat flows from one place to another (that is, a flow from node A to node B implies the material does not exist in A anymore). Opinions, by contrast, do not vanish after propagation (that is, A's inclination towards a candidate does not disappear even after communicating his or her opinions to B). Therefore we add a third phase to include a fusion model that integrate a individual's own opinions $S(t)$ with the opinion $S_n(t)$ gathered from its neighbors.

In the fusion phase, we introduce a parameter for each individual: the susceptible ratio, a real number $\epsilon \in [0, 1]$ that represents how easily a individual can be affected by others. Given a susceptibility parameter for each individual, we can then create a susceptible matrix E, an $n \times n$ diagonal matrix with the ϵ value of each individual in the diagonal. If E equals to the identity matrix I, which would imply all individuals are equally and highly susceptible to one another, then $S(t+1)$ should be equivalent to its neighbors' opinion $S_n(t)$. On the opposite side, if E equals to the zero matrix, implying all individuals are impervious to one another, then $S(t+1)$ should be identical to $S(t)$. Thus, after one iteration of propagation, the preference score matrix can be represented as

$$S(t+1) = (I-E)S(t) + ENFS(t) = ((I-E)+ENF)S(t).$$

Note that we assume that E does not change over time, and neither does F (which is only dependent upon topology). Interestingly, at first glance one might assume that N changes iteratively, it actually does not. Because the sum of each column in F equals 1 and the scores are always normalized for all candidates, it is not

hard to prove that

$$N_{ij} = \begin{cases} (\sum_{j=1}^{n} F_{i,j})^{-1} & \text{when } i = j \\ 0 & \text{otherwise} \end{cases}$$

which depends only on F. Therefore, we can write $S(t+1)$ as $\mathcal{X}S(t)$ where $\mathcal{X}$ is a time-independent matrix. This becomes an important feature for the proof of convergence in the next section.

Above concludes one iteration of propagation. In the next iteration, $S(t+1)$ becomes the initial preference scores for the individuals and the same process can be executed to obtain another round of propagation results $S(t+2)$. Below is the algorithm for our model.

Algorithm 1 Preference Propagation Model

R : iteration number; P : initial preference profiles
E : susceptible matrix; F : forwarding matrix
N : normalization matrix
$S(0) = PreferenceToScore(P)$
for $t = 0$ **to** R **do**
  $S_p(t) = FS(t)$
  $S_n(t) = NS_p(t)$
  $S(t+1) = ES(t) + (I - E)S_n(t)$
end for
return $S(R)$

2.2 Proof of Convergence and Consensus In this section, we show the convergent property of our proposed scheme. The score matrix becomes invariant after a sufficient number of propagations. Moreover, we show that given certain conditions all rows in the converged score matrix are identical. In the other words, a consensus within a community will eventually be reached through information propagations in our model.

Let $\mathcal{X}$ denote the overall preference propagation operation of all three phases explicitly laid down in the previous section,

$$S(t+1) = \mathcal{X}S(t) = [(I - E) + ENF]S(t)$$

To provide intuition for the forthcoming deductions and to borrow results of the properties of $\mathcal{X}$ from section 2.1, we start by pointing out the similarities as well as differences between $\mathcal{X}$ and the PageRank matrix $\mathcal{G}$. First, the entity $\mathcal{X}$ acting on $S(t)$, is actually a matrix consisting of the vectors of probabilities instead of a simple vector of probabilities. As a result, the columns of $\mathcal{X}$ do not add up to 1 (only the rows do) and therefore it is not a stochastic matrix. Furthermore, a social personal relationship network is intrinsically more localized compared to the World Wide Web, and as such, the favorable positive definite property enjoyed by $\mathcal{G}$ does not necessarily hold for $\mathcal{S}$. That said, these complexitie, while no doubt complicates the theoretical treatment of our algorithm, are in fact a natural manifestation of the increased richness of our target of research in hand — social networks.

We start our deduction of the convergence of $\mathcal{X}$ by enlisting the Perron-Frobenius theorem [14] which states that an irreducible, acyclic matrix has a single eigenvalue that is strictly larger than the others. Under the assumption that the graph being induced by $\mathcal{X}$, $G_{\mathcal{X}}$ is strongly connected and that the weights matrix E have entries smaller than one but not all zeros, $\mathcal{X}$ is irreducible and acyclic, and thus applies to the Perron-Frobenius theorem. We denote the dominant real positive eigenvalue of $\mathcal{X}$ as r. Armed with this fact, we are able to transform $\mathcal{X}$ into its Jordan canonical form

$$\mathcal{X} = P^{-1}J_{\mathcal{X}}P, J_{\mathcal{X}} = \begin{pmatrix} J_{\mathcal{X}_1} & 0 & \dots \\ 0 & J_{\mathcal{X}_2} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix},$$

by which the leading block $J_{\mathcal{X}_1}$ is a 1×1 matrix $[r]$, and other $J_{\mathcal{X}_i}$s correspond to their strictly smaller eigenvalues $\lambda_{\mathcal{X}_i}$. Since by the rules of matrix multiplication, the effect of $\mathcal{X}$ on $S(t)$ can be analyzed one by one with respect to $S(t)$'s column vectors without loss of generality, we will proceed on with our proof of $S(t)$'s convergence by concentrating on $S(t)$'s column vectors which we denote by lower case $s(t)$. Decomposing $s(0)$ into the sum of $\mathcal{X}$'s eigenvectors, $c_1v_1 + c_2v_2 + \dots$, we obtain the general form of the time evolution of $s(t)$,

$$s(t) = J_{\mathcal{X}}^t(c_1v_1 + c_2v_2 + \dots) = r^t(c_1v_1 + b_t),$$

where

$$\begin{aligned} ||b_t|| &= \tfrac{1}{r^t}||J_{\mathcal{X}_2}^t\, c_2v_2 + \dots|| \\ &\leq \textstyle\sum_{i=2}^{|V|} \left(\frac{|\lambda x_i|}{r}\right)^t ||c_iv_i|| \to 0, \text{as } t \to \infty \end{aligned}$$

The above shows that $||b_t||$ converges to zero when t is large, and therefore $S(t)$ converges to $r^t(c_1v_1)$. To get an intuition for the speed of this convergence, we turn to a special case where the susceptible ratios are identical, that is E is a scalar ϵ. In this case, we apply the Perron-Frobenius theorem again on NF, and we again obtain NF's Jordan form

$$NF = P^{-1}J_{NF}P, J_{NF} = \begin{pmatrix} J_{NF_1} & 0 & \dots \\ 0 & J_{NF_2} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

However, note that since it needs not be acyclic, be strictly larger than the other. Now, using this basis we find that $\mathcal{X}$ equals to

$$\epsilon \begin{pmatrix} \ddots & \ddots & & & 0 \\ & \frac{(1-\epsilon+\epsilon\lambda_{NF_i})}{\epsilon} & 1 & & \\ & & \frac{(1-\epsilon+\epsilon\lambda_{NF_i})}{\epsilon} & 1 & \\ & & & \frac{(1-\epsilon+\epsilon\lambda_{NF_i})}{\epsilon} & \ddots \\ 0 & & & & \ddots \end{pmatrix}$$

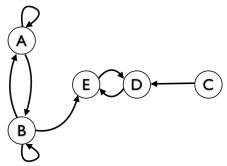


Figure 1: Nodes A, B form an opinion leader SCC, while node C by itself is another opinion leader SCC. Nodes E and D form an opinion follower SCC.

Since a Jordan canonical form is unique, we obtain $\lambda_{\mathcal{X}_i} = (1 - \epsilon + \epsilon\lambda_{NF_i})/\epsilon$. From this result, we confirm that when $\epsilon = 0$, $\mathcal{X}$ degenerates to the trivial diagonal case; and that as ϵ approaches 1, the rate of convergence is geometrically proportional to ϵ/r.

We are now one step away from the final proof of S's convergence. Recalling that $s(t) \to r^t c_1 v_1$, once $r \leq 1$ is established, $S(t)$ converges. To prove this, we take advantage of the Collatz-Wielandt theorem which gives the following formula for r: $r = \max_{x \in N} f(x)$, where $f(x) = \min_{1 \leq i \leq n; x_i \neq 0} \frac{[\mathcal{X}x]}{x_i}$, and $N = \{x | x \geq 0 \text{ with } x \neq 0\}$.

We begin by asserting that the upper bound of $f(x)$ is 1. To prove this, we suppose the opposite holds, that means there exists x such that $f(x) = \min_{1 \leq i \leq n; x_i \neq 0} \frac{[\mathcal{X}x]}{x_i} = \alpha > 1$. This implies the following list of equations:

$$\begin{gathered} 1 < \alpha \leq \tfrac{1}{x_1}(\mathcal{X}_{11}x_1 + \mathcal{X}_{12}x_2 + \ldots + \mathcal{X}_{1n}x_n) \\ \vdots \\ 1 < \alpha \leq \tfrac{1}{x_n}(\mathcal{X}_{n1}x_1 + \mathcal{X}_{n2}x_2 + \ldots + \mathcal{X}_{nn}x_n) \end{gathered}$$

Note that $\sum_{j=1}^{n} \mathcal{X}_{ij} = 1, \forall i$. Thus, the above list of equations can be arranged into

$$\begin{gathered} \mathcal{X}_{12}(\tfrac{x_2}{x_1} - 1) + \mathcal{X}_{13}(\tfrac{x_3}{x_1} - 1) + \ldots + \mathcal{X}_{1n}(\tfrac{x_n}{x_1} - 1) > 0 \\ \vdots \\ \mathcal{X}_{n1}(\tfrac{x_1}{x_n} - 1) + \mathcal{X}_{n2}(\tfrac{x_2}{x_n} - 1) + \ldots + \mathcal{X}_{n(n-1)}(\tfrac{x_{n-1}}{x_n} - 1) > 0 \end{gathered}$$

However, by denoting i as the subscript that has $x_i = \max_{1 \leq j \leq n} x_j$ and remembering that $\mathcal{X}$ is a non-negative matrix, one of the above equations would not hold.

$$\mathcal{X}_{i1}(\frac{x_1}{x_i} - 1) + \mathcal{X}_{i1}(\frac{x_2}{x_i} - 1) + \ldots + \mathcal{X}_{in}(\frac{x_n}{x_i} - 1) > 0$$

This justifies the assertion that $f(x) \leq 1$. Combining this result with the observation that the trivial vector $(1, 1, \ldots)$ yields $f(x) = 1$, we conclude that $\max_{x \in N} f(x) = 1$. Therefore, $r = 1$, and $S(t)$ converges to $c_1 v_1$.

For networks that are not strongly connected we can always find the SCCs in linear time, and the problem reduces to the smaller "source SCCs" of the network since the matrices of all the other SCCs have a Perron root smaller than 1 and their elements eventually vanish. For the remaining source SCCs, since no vertices has susceptibility ratios equals 1, according to the above results they all converge. The net effect is exemplified by the stark difference between the individuals belonging to the various source SCCs and the rest. Whereas source SCC vertices will converge to their own respective common values, the others may converge to different values and act as followers in terms of aligning their own preferences to the weighted average of those belonging to the sources. Figure 1 gives an example of such phenomenon. Let the initial preference matrix of all the nodes in Figure 1 be

$$\begin{pmatrix} s_A & s_B & s_C & s_D & s_E \\ s'_A & s'_B & s'_C & s'_D & s'_E \\ s''_A & s''_B & s''_C & s''_D & s''_E \end{pmatrix} \begin{matrix} \text{candidate1} \\ \text{candidate2} \\ \text{candidate3} \end{matrix}$$

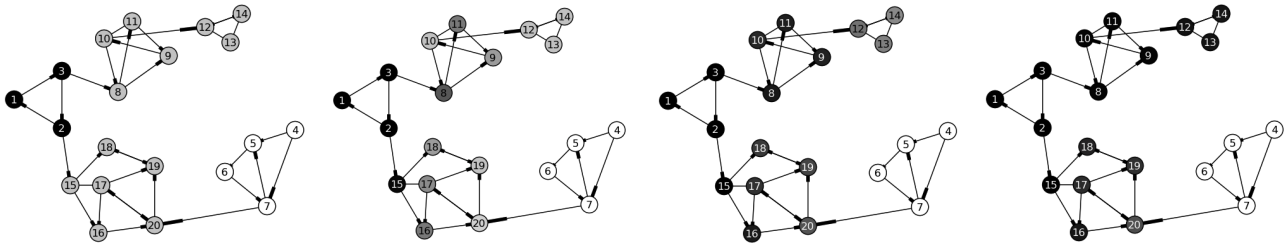
where each row in the preference matrix denotes each node's preference for candidate 1, 2, and 3 respectively. Then after infinite number of propagations, the preference matrix will become

$$\begin{pmatrix} s(\infty)_{AB} & s(\infty)_{AB} & s(\infty)_C & s(\infty)_D & s(\infty)_E \\ s(\infty)'_{AB} & s(\infty)'_{AB} & s(\infty)'_C & s(\infty)'_D & s(\infty)'_E \\ s(\infty)''_{AB} & s(\infty)''_{AB} & s(\infty)''_C & s(\infty)''_D & s(\infty)''_E \end{pmatrix}$$

in which the preferences of nodes A and B in Figure 1 for candidate 1 converges to the common value $s(\infty)_{AB}$; for candidate 2 converges to the common value $s(\infty)'_{AB}$; and for candidate 3 converges to the common value $s(\infty)''_{AB}$. However, for nodes D and E, given that the SCC composed by them $\{D, E\}$ is under the influences of both opinion leaders SCC $\{A, B\}$ and $\{C\}$, their eventual preferences instead of aligning themselves to a common value becomes a linear combination of the preferences of $\{A, B\}$ and $\{C\}$. The exact details of this combination depend on the structure of the network.

The preference propagation model simulates this unique behavior of people by projecting the preferences vector onto the leading uniform eigenvector denoting equilibrium. In addition, it also attempts to mimic the real world by distinguishing the opinion leaders from the followers. As with its real world counterpart, this process is solely determined by the initial preferences of every individual and the structure of the embedding social network.

Another example is shown in Figure 2, time evolution of preferences held by nodes in a social network, demonstrating the effects of opinion leaders creating their own consensus and passing it down to opinion followers in a cascading manner. We see that the opinion follower SCC composed by nodes 15 to 20 are colored with various shades of gray depending on their distance to the two opinion leader SCCs composed by nodes 1 to 3 and 4 to 7. We also observe that the preference of the opinion leader SCC 1 to 3 is first passed to the opinion



(a) round 0 (b) round 2 (c) round 10 (d) round 29

Figure 2: Time evolution of preferences held by nodes in a social network, demonstrating the effects of opinion leaders creating their own consensus and passing it down to opinion followers in a cascading manner.

follower SCC 8 to 11 (in propagation round 10), and then subsequently passed to the opinion follower SCC 12 to 14 through the efforts of the SCC 8 to 11 in a cascaded manner.

This simple example demonstrates that the strongly-connection source components form the opinion leader groups, while each follower node is affected by (i.e. linear combination) the opinions of its surrounding opinion leader groups. Our framework models the real-world observation about how each less-convinced personnel being affected by the mass opinions he or she encountered.

2.3 Comparison with other models We here discuss what are the most salient characteristics of a successful social model based on common observations and beliefs, in an attempt to contrast the most distinguishing features of our model with the other previously proposed frameworks.

Hyper Dimension Media. Since a personal preference describes the order of preference of all possible candidates, the media in an ideal model should be represented as ordered lists instead of a single value. Most of the propagation models such as Linear Threshold Model, Cascade Independent Model, or Diffusion Rank, unfortunately, only handle binary or real value in propagation.

Topology Dependent and Input Dependent. The word-of-mouth is the main strategy for a person to affect others. The real-world process of guiding friends toward the adoption of self preference goes mutually and simultaneously. To state such phenomenon, the outgoing persuasions of a person should ideally become a combination of self-preference and the incoming preferences. An ideal model should both take into account of network structure and initial personal preference. Moreover, we would like a model's way of incorporating these two factors to be as natural as possible, instead of relying on ad hoc stopping designs or simply restricting the number of times nodes or individuals interact.

Deterministic Convergence. Of course an ideal model should converge or end eventually, or else it would be difficult for the modeler to interpret the results. As far as we know, there are currently two kinds of designs to achieve such a convergence. The first one, such as LT model and IC model, attaches a binary status to each node in a network to determine whether it is visited. The *inactive* status means the node is not yet visited while the *active* status means the node is visited. With such design, preference propagation to inactive nodes can be easily monitored. Moreover, the propagation converges in such model when none of the existing node can change the status anymore.

Following the success of the PageRank algorithm, the second popular approach is building the convergence mechanism into a model inherently, so that after sufficient iterations the model converges and produces a definite result.

To make results easily analyzable, convergent models that can generate repeatable results given both the same initial preferences of nodes and network structure are preferred.

Consensus. The problem of reaching a consensus among agents has been studied since around 1970 [3, 17] with simulation models such as the voter model [11]. Mossels et al. gave a theoretical prove that the consensus could be reached with the voter model. Thus an ideal model should be able to reflect specific common traits. In particular, we observe that one such universal trait is people in the same community (i.e. SCC) have the tendency to align their preference after sufficient exchanges. This translates into the fact that an ideal model should contain some kind of homogeneity inside a group.

To see how our model and other proposed frameworks capture the above characteristics of real world social interactions, we conducted several experiments and recorded their results in Table 2 below for ease of comparison. We particularly chose models that are most representative in their own stance, namely the Linear Threshold model, Independent Cascade model, PageR-

Table 2: Comparison of models on the abilities to capture characteristics of real world social networks interactions.

	Convergence	Repeatability of final state	Consensus	Input dependent	Media space
Proposed Model	✓	✓	✓(if SCC)	✓	R^k
LT Model	✓	✓		✓	boolean
IC Model	✓			✓	boolean
PageRank	✓	✓			R
DiffusionRank	✓	✓		✓	R

ank model, and DiffusionRank model for comparison. Note that since the propagating media in these models are not a vector of preference, we made the following enhancements for each of them to handle such cases. For the LT and IC models, we assume that each vertex initially held approval for its top k preferred candidates (non approval for the others), and thus for every candidate, we get a list of seeds as input into the LT and IC models. We then ran the model separately on each candidate, garner their results, and normalize them to form the final preference of each vertex. For the PageRank and DiffusionRank models, given that they can take real values as inputs, we simply executed these models separately for each candidate in the preference list, and then integrate the results to be a vector of real numbers. Lastly, as shown in Table 2, we see that our model is the only model that operates directly on a list of preference, whereas other models work restrictively on single boolean or real values, and have to be executed separately to obtain a joint preference, which fail to consider the correlation of the preference score among candidates.

To see whether these models both converge and produce repeatable results, for each model we ran three identical experiments with the same initial preferences and network. We note that all models gave convergent results. Besides, since the IC model carries a random component, it does not deliver repeatable final preference results.

To examine whether these models can give a kind of consensus to nodes that belong to a strongly connected network, we ran all models on a strongly connected graph until they naturally stops or converges. It turns out that except our model, none showed signs of reaching consensus among the final output preferences. Note that our model does not produce consensus given non-SCC components.

To see whether these models take into account of the initial preferences held by nodes, we fed all models with six different initial preferences and see whether they give six different results. It is not surprising that the PageRank model returns identical results regardless of the input, indicating that it takes into account of only the structure of the network but ignoring the initial preferences held by each node or individual. In conclusion, our model is the only framework that supports all five criteria set by observations from real world social networks.

3 Experiment

To evaluate the performance, we compare our models with some well-known diffusion models such as Linear Threshold, Independent Cascade, and DiffusionRank in the experiment. Ideally, we will examine whether all the mentioned algorithms including ours can capture the preference transition in social networks to a certain extent. Conducting such validation, some information is vitally required such as the network structure, and the preference for nodes over time.

3.1 Preference Data In scientific research papers, citations implicitly reveal the research interests of authors. In other words, we believe that the acts such as citing or submitting to the journals or the conferences would reveal the authors' interests. By utilizing this fact, we can infer the researchers' preference from their corresponding top frequently-cited conferences and journals. Thus, it is possible to express the research interests as the preference on conferences and journals. Furthermore, we have realized that the collaborations with others, one may gradually change his own preference. It is particularly correct for advisor-student relationship since the advisors and students usually affect each others' research interests and directions. We have designed an experiment to model how researchers' preferences can be affected by the collaborators in social network.

We use KDD Cup 2003 ArXiv HEP-TH (High Energy Physics - Theory) citation network [4] with the corresponding paper meta information as our evaluation dataset. This citation dataset spreads over 12 years from 1992 to 2003. We choose the top 16 journals that possess most papers as the candidates to construct the preference lists. We construct the yearly preference lists based on the citation count of the corresponding journals within a year in our case. Note that we prefer using the citations rather than the publications of authors because the publications imply not only preference but also capability. To fairly present the interests, we use the citations. In addition, we construct a collaborative network from this dataset as the underlying social preference diffusion backbone. To easily perceive the changes in interests, we remove the authors who had fewer than 5 publications in the dataset, which results

in a network with 2683 nodes.

3.2 Model Comparison Since we already have all the required information including network structure and preference transition, the next step is to study which diffusion model predicts the preference transition better. We assume a good diffusion model could capture the progression the authors' research interests through collaborations. To do so, we initially set up the node preference according to the actual data in year x, and then compare the predicting results with the actual preference in year $x + k$. Following issues are noted in the experiment:

Hyper Dimension Media. To represent the order in preference toward all candidates, the media in an ideal model ought to be an ordered list instead of a single value. Nonetheless, most well-known diffusion models such as LT, IC, and DiffusionRank, only treat the media as boolean or real number. For comparison, we exploit these models in our problem by executing them independently for each candidate. We evaluate the candidate rank based on each independent diffusion result.

Determinism of the Final State. Except the IC model, outcome of all the models mentioned above is deterministic. Because the parameter called diffusion probability in IC model is a nondeterministic factor, we execute the experiment 20 times and average the results.

Initialization. Because the media in LT and IC model are not native for hyper dimension, we singly process the propagation for each candidate. That means, in our experiment, the active mode of top 1% authors to a specific publisher are initially set active in LT and IC models while the rest publishers are set inactive. We further set the diffusion probability of each edge as $\frac{1}{N}$, where N is the degree of its source node in IC model. In LT model, we assign links with identical weight, and nodes with same threshold. The parameters in LT and IC are then tuned to find the optimal outcome. The propagation process is executed multiple times with different thresholds and the performance are averaged. For DiffusionRank model, we use the parameter settings suggested by the authors of [18].

3.3 Experiment Result Diffusion models are evaluated by comparing their predictions about preference in 1997, 1998, and 1999 while using the real preference during the period since 1993 to 1996 as initial status. To measure the similarity between predicting and real results, we adopt the Kendall's tau coefficient [9] and the Jaccard coefficient. We individually measure the similarity for each author, each node in the network, and then average them as a performance indicator. Because Kendall's tau coefficient is not well-defined with tie scores, we manually set Kendall's tau score as 0 when there is a tie on all 16 publishers. Furthermore, we calculate the Jaccard coefficient performs on top 3 highest scored publishers.

	Kendall's Tau			Top 3 Jaccard		
year	1997	1998	1999	1997	1998	1999
IC	0.007	0.012	0.015	0.011	0.014	0.015
LT	0.172	0.167	0.167	0.171	0.195	0.212
DiffusionRank	0.221	0.181	0.160	0.216	0.222	0.213
proposed(0.00)	0.240	0.204	0.178	0.242	0.243	0.225
proposed(0.25)	0.243	0.206	0.180	0.248	0.244	0.226
proposed(0.50)	0.243	0.206	0.180	0.247	0.243	0.227
proposed(0.75)	0.243	0.206	0.180	0.246	0.243	0.226
proposed(1.00)	0.230	0.190	0.163	0.204	0.179	0.156

Table 3: Compare the result after one round for each model with the ground-truth of year 1997, 1998, and 1999.

Firstly, for the sake of knowing the correspondence between the extent of changes in iterations and in years, we execute one-iteration propagation in each model, and then compare the results with the ground truth in 1997, 1998, and 1999 respectively. We also try different susceptible ratio ϵ in our model, as $\epsilon = 1.0$ implies the authors stick to their own preferences without considering the effect from the neighbors. Table 3 shows the results, we find that it is quiet suitable to take one iteration as a period of a year. The results demonstrate that our model consistently outperforms the 2nd best model DiffusionRank, regardless which susceptible ratio is using as long as it is not 1.0.

Secondly, we execute the diffusion algorithms for multiple rounds, and compare it with the ground truth of year 1997-1999. Table 4 shows the average of the scores for 1997, 1998, and 1999. Note that LT and IC model stop when there is no possible activation (regarded as one round), which implies that authors are not affected by their neighbors after the first round completes. Table 3 and 4 additionally show that the impervious preferences ($\epsilon = 0$) reach a performance similar to the best result, which might reveal the slowly changing nature. Nevertheless, the results show that our model can faithfully capture the trait of the social influence even the authors' interests change slowly.

4 Conclusion

Analysing the effect of social networks upon group decisions outcomes is a difficult problem because it is both costly and time consuming to perform user studies to collect people's private preferences. Indeed, it is the change of preferences through social propagation in particular that we care most about, and to our knowledge this is the first ever study that provides not only theoretical analysis but the empirical justification of this problem. This study provides an example of how to perform such research with limited data through exploiting al-

	Kendall's Tau					Top 3 Jaccard				
round	1	2	3	4	5	1	2	3	4	5
Independent Cascade	0.011	0.011	0.011	0.011	0.011	0.013	0.013	0.013	0.013	0.013
Linear Threshold	0.168	0.168	0.168	0.168	0.168	0.192	0.192	0.192	0.192	0.192
DiffusionRank	0.186	0.186	0.186	0.186	0.186	0.217	0.217	0.217	0.217	0.217
proposed(0.00)	0.208	0.209	0.207	0.206	0.205	0.238	0.240	0.237	0.236	0.234
proposed(0.25)	**0.210**	0.209	0.208	0.207	0.207	0.241	0.241	0.240	0.239	0.238
proposed(0.50)	0.209	0.210	0.209	0.209	0.208	0.240	0.242	0.242	0.241	0.240
proposed(0.75)	0.209	0.209	0.209	0.209	0.209	0.239	0.241	0.242	**0.242**	0.241
proposed(1.00)	0.194	0.194	0.194	0.194	0.194	0.179	0.179	0.179	0.179	0.179

Table 4: Consider the result after $k \times R$ rounds for each model, and compare it with the ground-truth of year $1996 + k$. The table shows the average of the similarity scores for 1997, 1998, and 1999.

gorithm and model design, theoretical justification, and computer simulation.

Another significant contribution of our work is that we provide an alternative evaluation plan and data to verify a preference propagation model. Acknowledging the lack of real-world data to evaluate how the voter's preference can change through social diffusion, we have come up with a novel idea to identify a publicly available bibliography dataset to evaluate how researchers gradually change their research fields according to the influence of their collaborators. Our evaluation plan opens a new possibility that allows researchers working on preference diffusion problems to be able to evaluate their models without having to identify a highly private voter preference dataset.

5 Acknowledgement

This work was supported by National Science Council, National Taiwan University and Intel Corporation under Grants NSC101-2911-I-002-001, NSC101-2628-E-002-028-MY2 and NTU102R7501.

References

[1] J. J. Bartholdi, C. A. Tovey, and M. A. Trick. How hard is it to control an election. In *Mathematical and Computer Modeling*, pages 27–40, 1992.

[2] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *SIGKDD*, pages 199–208. ACM, 2009.

[3] M. H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):pp. 118–121, 1974.

[4] J. Gehrke, P. Ginsparg, and J. Kleinberg. Overview of the 2003 KDD Cup. *SIGKDD Explor. Newsl.*, 5(2):149–151, Dec. 2003.

[5] A. Gibbard. Manipulation of voting schemes: A general result. *Econometrica*, 41(4):587–601, 1973.

[6] M. Granovetter. Threshold models of collective behavior. In *American Journal of Sociology83(6)*, pages 1420–1443, 1978.

[7] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. WWW '04, pages 491–501. ACM, 2004.

[8] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *ACM KDD*, pages 137–146, 2003.

[9] M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2), 1938.

[10] J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs: Patterns and a model. Technical report, 2006.

[11] T. Liggett. *Interacting particle systems*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, 1985.

[12] F. C. Liu. Modeling political individuals using the agent-based approach: A preliminary case study on political experts and their limited influence within communication networks. *Journal of Computers*, 19(4):8–19, 2009.

[13] H. Ma, H. Yang, M. R. Lyu, and King. I.: Mining social networks using heat diffusion processes for marketing candidates selection. In *CIKM*, pages 233–242, 2008.

[14] C. D. Meyer, editor. *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

[15] K. Saito, M. Kimura, K. Ohara, and H. Motoda. Selecting information diffusion models over social networks for behavioral analysis. volume 6323 of *Lecture Notes in Computer Science*, pages 180–195. 2010.

[16] M. A. Satterthwaite. Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2):187–217, April 1975.

[17] R. L. Winkler. The consensus of subjective probability distributions. *Management Science*, 15(2):B61–B75, 1968.

[18] H. Yang, I. King, and M. R. Lyu. Diffusionrank: A possible penicillin for web spamming. In *SIGIR 2007*, 2007.

[19] S. Yoo, Y. Yang, F. Lin, and I.-C. Moon. Mining social networks for personalized email prioritization. KDD '09, pages 967–976. ACM, 2009.