Information Flow Modeling based on Diffusion Rate for Prediction and Ranking

Xiaodan Song, Yun Chi, Koji Hino, Belle L. Tseng
NEC Laboratories America, 10080 N. Wolfe Road, SW3-350, Cupertino, CA 95014, USA
{xiaodan, ychi, hino, belle}@sv.nec-labs.com

ABSTRACT

Information flows in a network where individuals influence each other. The diffusion rate captures how efficiently the information can diffuse among the users in the network. We propose an information flow model that leverages diffusion rates for: (1) prediction – identify where information should flow to, and (2) ranking – identify who will most quickly receive the information. For prediction, we measure how likely information will propagate from a specific sender to a specific receiver during a certain time period. Accordingly a rate-based recommendation algorithm is proposed that predicts who will most likely receive the information during a limited time period. For ranking, we estimate the expected time for information diffusion to reach a specific user in a network. Subsequently, a DiffusionRank algorithm is proposed that ranks users based on how quickly information will flow to them. Experiments on two datasets demonstrate the effectiveness of the proposed algorithms to both improve the recommendation performance and rank users by the efficiency of information flow.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval – *Information Filtering*; J.4 **[Computer Applications]**: Social and Behavioral Sciences – *Economics*

General Terms

Algorithms, Experimentation

Keywords

Information Flow, Social Influence, Diffusion of Innovation, Continuous-Time Markov Chain, Recommendation, Collaborative Filtering, Web Ranking

1. INTRODUCTION

People constantly influence each other in all facets of life. *Social influence* describes the phenomenon by which the behavior of an individual is *directly* or *indirectly* affected by the thoughts, feelings, and actions of others in a population [1][21]. Such influence is present when people recommend products or services to one another. Word-of-mouth communication is a particular type of informational social influence, and plays an important role in shaping the attitudes and behaviors of consumers.

When one needs to choose among various options where he/she has no experience, one will often rely on the opinions of others with such experiences. However, when there are thousands or millions of options, like on the Web, it becomes practically impossible for an individual to identify reliable experts that can give advice about each of the options. Collaborative Filtering was proposed to automate the process of "word-of-mouth" [2] by leveraging like-minded users' opinions. It infers the interests/preferences of an individual based on the interests/preferences of people with similar tastes [3][4]. In Collaborative Filtering, *similarity* - which is a symmetric relationship between users - plays a central role. However, it is in fact the *inter-personal influence* - an *asymmetric* relationship - that most directly and effectively supports the automation of the word-of-mouth process. For instance, asymmetric relationships such as employer-to-employee, teacher-to-student, and physician-to-patient have a much stronger influence on decisions than their reverse relationships. We recently proposed a novel asymmetric recommendation algorithm based on such asymmetric inter-personal relationships [5]. The algorithm leverages a user's explicit or implicit influence over other users, instead of the mere symmetric similarity of users' interests in Collaborative Filtering.

People adopt new technologies at different times. The idea of modeling the adoption behaviors of a group of people as a flow process comes from Rogers' "Diffusion of Innovation" theory [6], in which the adoption curve classifies adopters of innovations into five categories based on the fact that certain individuals are inevitably more open to adoption than others in a population. The innovation can be an idea, a practice, or an object. The five adopter categories, (1) innovators, (2) early adopters, (3) early majority, (4) late majority, and (5) laggards, follow a standard deviation curve. Innovators and early adopters are usually the social leaders. Early majority, late majority, and laggards are the followers. The perceived novelty of the idea by an individual determines his/her reaction to it. As a result, diffusion is the process by which an innovation is communicated through certain channels over time among the members of a population. The information can be imagined flowing from social leaders to followers. The behavior of followers is directly or indirectly influenced by the actions of social leaders. In this paper, we use terms "influence", "information flow", and "information diffusion" interchangeably to represent the same concept.

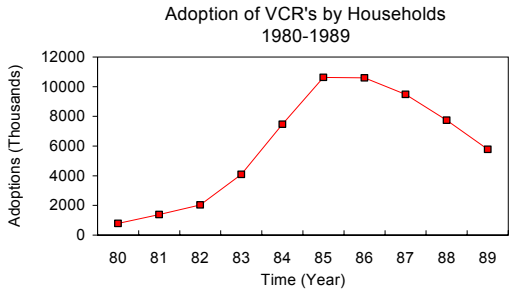
Figure 1 (a) provides an intuitive example of the diffusion of innovation theory on the adoption of VCRs in the 1980s [7]. As illustrated in Figure 1(a), when VCRs enter the market, only around one million users adopt them. From 1980 to 1986, the number of adoptions keeps increasing, and finally declines after 1986. Similar diffusion patterns exist on other products and technologies. We may remember how the cell phones are adopted in our family, as shown in Figure 1(b). One month after the cool neighbor (u_1) purchases a cell phone, the kid (u_2) notices it and persuades the family to get him one to contact the friends at school. Then the father(u_3) also finds that he needs it for work and decides to get another one, following is the mother (u_4). Finally, it takes a long time for the grandmother (u_5) to accept the cell phone technology and purchase one.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). Distribution of these papers is limited to classroom use, and personal use by others.
WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

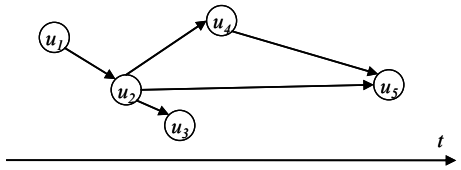
Diffusion has a time dimension, which Rogers describes as the "innovation-decision" process [6]. The influence of central users tends to spread more rapidly throughout a network than the influence of peripheral users does. Also, it usually takes less time to reach central users than peripheral users when information flows in the network. Information flows in different rates through the network. Our previous model [5] ignores the fact that the network is not homogeneous in terms of the diffusion rate.

Let us continue the previous example with users u_1 to u_5, representing the neighbor, the kid, the father, the mother, and the grandmother respectively as illustrated in Figure 1(b). Although the father, the mother, and the grandmother will follow whatever the kid adopts (i.e., the kid influences the father, the mother, and the grandmother in the same amount), the father always follows the kid within one day, while it will take more than a year for the grandmother to follow up. Information flows from the kid to the father much faster than to the grandmother. Given the kid adopts one item at time t, it is more likely that the father will adopt it at time $t+1$ than the grandmother, although the grandmother may eventually be interested in the item and adopt it at a much later stage. Therefore, it attains higher successful rate to recommend the item to the father than to the grandmother at time $t+1$.

On the other hand, let us look at the problem of ranking these five users to find out the most efficient receiver of information. As what we have analyzed, the father should be the most efficient receiver in the network. However, by using existing ranking algorithms such as PageRank [9] or HITS [10], which do not take the diffusion rate into account, we can not find the correct answer, and instead, the grandmother is ranked as the most important receiver in the network.



(a) Adoption of VCR's by households (data is obtained from [7]).



(b) An example of adoption. Edge indicates the influence in the network. Only the most direct and efficient influence is illustrated.

Figure 1: Intuitive examples of diffusion of innovation theory on the adoption

In this paper, we propose an information flow model with inter-personal diffusion rate taken into account to measure how *efficiently* information diffuses from one user to another in the information flow network. The algorithm is based on *Continuous-Time Markov Chain*, in which we model both the probability (how likely) and the rate (how fast) for the information to flow from one user to others. We propose a prediction algorithm and a ranking algorithm based on this inter-personal diffusion rate based information flow model. We demonstrate that by taking the inter-personal diffusion rate into account, we can improve the recommendation performance and rank users by how efficiently they influence others.

The primary contributions of this paper are three-fold. First we propose a novel information flow model to leverage inter-personal diffusion rate based on Continuous-Time Markov Chain. Second, the rate-based information flow model is formulated to provide recommendations to users. Third, a DiffusionRank algorithm is defined to measure how efficiently information can flow to certain users.

The rest of the paper is organized as follows. We review related work in Section 2. Section 3 provides the problem formulation of leveraging rate-based information flow for prediction and ranking. In Section 4, we propose the information flow model based on Continuous-Time Markov Chain. In Section 5, we demonstrate the experimental results by conducting our rate-based information flow model on improving recommendation performance and ranking effective influential users. Finally, conclusions and future work are given in Section 6.

2. RELATED WORK

In this section, we review the diffusion of innovation theory, the related work on recommendation algorithms, and the Markov chain based website ranking algorithms.

2.1 Diffusion of Innovation Theory

In [6], Everett Rogers describes the different stages of product adoption and indicates that the spread of a new technology depends mainly on two factors; *innovation* or *imitation*. Innovators are driven by their desire to try innovations; in contrast, imitators are primarily influenced by the behaviors of their peers. The Bass model [15] quantifies the concept of introduction of products/technologies by estimating the introduction (innovation) and acceptance (imitation) rate variables. The model is widely used in market analysis and demand forecasting of innovation diffusion in various areas.

The Bass model characterizes the spread of a new product and technology in a market by

$$N(t) = N(t-1) + p\left(m - N(t-1)\right) + q\frac{N(t-1)}{m}\left(m - N(t-1)\right) \quad (1)$$

where $N(t)$ is the cumulative number of adopters by time t; the parameter m is the market potential, indicating the total number of people who will eventually adopt the item; the coefficient p is called the coefficient of innovation, indicating the external influence or advertising effect; the coefficient q is called the coefficient of imitation, indicating internal influence or word-of-mouth effect.

However, the Bass model ignores the network structure, which could significantly influence the diffusion process, and only takes the global diffusion rate into consideration.

Identifying how information is propagated in a network is important in various applications. Network-based marketing refers to a collection of marketing techniques that take advantage of links between consumers to increase sales [16]. [17-19] attempt to model influence among consumers and understand how this influence propagates in the networks. The problem they address is: suppose that we have data on a potential network of consumers with estimation of the extent to which individuals influence one

another, if we are given a new product and a marketing budget, how can we maximize the adoption of the new product through customers? The threshold model and the cascade model have been considered in [19]. Information propagation through blogosphere is also recently studied [20]. These models ignore the inter-personal diffusion rate, which describes how rapidly the information propagates in the network.

2.2 Recommendation Algorithms

Various Collaborative Filtering algorithms have been designed to identify users of similar interests [3]. The similarity metrics include cosine distance, correlation, and mean-squared difference. [8] utilizes the average commute time in a network to calculate the distance of a pair of users.

However, as we demonstrate in [5], symmetric similarity is not as direct and effective as *asymmetric inter-personal influence* between people for recommendation. We observe people intentionally or unintentionally influence and inspire each other, thus creating an interest in retrieving or getting a specific kind of information or product. Therefore, we model the information adoption behaviors of a group of people as an information flow process. This process captures the amount and direction in which information is propagated in a network to predict where the information may flow to. Effective recommendations are generated as the result.

However, our previous information flow model ignores the diffusion rate, which describes how efficiently information flows from one user to another in the network.

2.3 Ranking Algorithms

Markov Chain-based models have been proven successful in ranking web pages ([9-11]). The PageRank algorithm [9][10] computes the "importance" of web pages through a stochastic ergodic Markov transition matrix, which is constructed from all hyperlinks between web pages. The HITS algorithm [11] forms both an authority matrix and a hub matrix from the hyperlink adjacency matrix, rather than one Markov chain. As a result, HITS returns both authority and hub scores for each web page.

The research discussed above has focused on using static properties, such as the connectivity of the nodes in a network and the average node distances, to represent the complex structure. However, networks evolve over time. Dynamic factors in the context of the web are essential and could not be ignored. Based on the fact that users are not only interested in the pages with high authority scores, but also the recent information, [12] proposes T-Rank to take into account the temporal aspects such as freshness and update activity of pages and links when computing the importance of a page. [13] proposes a Markov centrality based on the mean first-passage time to measure the relative importance in networks. IRank [14] utilizes the timing ordering of the blogs to infer implicit links between blogs and rank the blogs according to those implicit links.

However, to the best of our knowledge, the diffusion rate, which shows how long it takes for a node in the network, such as a webpage or a user, to be aware of other nodes and make an explicit or implicit link, is ignored in previous research.

3. PROBLEM FORMULATION

The central problem we address in this paper is the efficiency of information flow. On the inter-personal level, we measure how likely information goes from a specific sender to a specific receiver during a limited time period. On the individual level, we estimate the expected rate that information diffuses to a particular node starting from an arbitrary node in the network. Thus, prediction of users' preferences of information, and ranking users by the efficiency of information flow can be formulated as follows.

Assume we have a network $G(n,w,\tau)$ containing a set of nodes n with a size of N, where edges between nodes represent the information flow paths, w denotes the weights on the edges to represent the amount of information flow from one node to another, and τ denotes the time delay on information flow paths.

3.1 Recommendation

For the problem of prediction of users' preferences of information, the question we address is how likely for information to go from a specific sender to a specific receiver within a certain time period. Given time t_0 and $\{u,z\} \subset n$, where u represents those who have adopted the item, and z represents those who have not, we calculate the likelihood that users z will follow users u to adopt the item before time t. We denote this likelihood as $L(z \mid u,t_0,t)$.

3.2 Ranking

For ranking problem, the question we address is what the expected time is for a user to receive information in a network. We calculate on average, how efficient for the information flowing from other users to a particular user i, and rank users based on the diffusion "efficiency", which is denoted as $R(i)$.

4. INFORMATION FLOW MODEL BY LEVERAGING DIFFUSION RATE

In this section, we review some related background on Continuous-Time Markov Chain. We then propose our rate-based information flow model based on the foundation of Continuous-Time Markov Chain. Afterwards we discuss how to utilize the model for generating recommendations and ranking users.

4.1 Continuous-Time Markov Chain

Definition 1: A Continuous-Time Markov Chain (CTMC) is a continuous time stochastic process $\{X(t), t \geq 0\}$ s.t. $\forall s,t \geq 0$, and $\forall i,j,x(h)$.

$$\begin{aligned} &P\{X(t+s)=j \mid X(t)=i, X(h)=x(h), 0 \leq h \leq t\} \\ &= P\{X(t+s)=j \mid X(t)=i\} \end{aligned} \quad (2)$$

A Continuous-Time Markov Chain satisfies the Markov property and takes value from a discrete state space. The Markov property states that at any times $t+s>t>0$, the conditional probability distribution of the process at time $t+s$ given the whole history of the process up to and including time t, depends only on the state of the process at time t [28]. In this paper, we assume the transition probabilities are independent from the initial time t, which means the chain is time-homogeneous and we denote $P_{ij}(s)$ as the transition probability from i to j over s time period.

Definition 2: Define the transition rate matrix as

$$\mathbf{Q} = \begin{pmatrix} q_{0,0} & q_{0,1} & \cdots \\ q_{1,0} & q_{1,1} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \quad (3)$$

where

$$q_{ij} = \lim_{\Delta t \to 0} \frac{P\{X_{t+\Delta t} = j \mid X_t = i\}}{\Delta t} = \lim_{\Delta t \to 0} \frac{P_{ij}(\Delta t)}{\Delta t} \qquad (i \neq j) \quad (4)$$

as the probability per time unit that the CTMC makes a transition from state *i* to state *j* or the transition rate. Thus the total transition rate out of state *i*, which we call out-state rate, is

$$q_i = \sum_{j \neq i} q_{i,j} \quad (5)$$

Define $q_{i,i} = -q_i$, which means when the chain leaves state *i* with rate q_i, it must enter some other states *j*'s, then

$$\mathbf{Q} = \begin{pmatrix} q_{0,0} & q_{0,1} & \cdots \\ q_{1,0} & q_{1,1} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} -q_0 & q_{0,1} & \cdots \\ q_{1,0} & -q_1 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \quad (6)$$

Definition 3: Define the time until the CTMC makes a transition and leaves state *i*, given that the CTMC is currently in state *i*, as the state staying time of the chain in state *i*, T_i.

$$T_i := \inf\{t : X_t \neq i \mid X_0 = i\} \quad (7)$$

where inf denotes inferior limit. T_i is exponentially distributed with rate q_i. When the stochastic process leaves state *i*, it will next enter state *j* with probability P_{ij}, which is independent of the time spent at state *i*, and satisfies

$$\begin{cases} \sum_{j \neq i} P_{ij} = 1 \\ P_{ii} = 0 \end{cases} \quad (8)$$

Also we have

$$P_{ij} = \frac{q_{ij}}{q_i} \qquad (i \neq j) \quad (9)$$

In a summary, in a CTMC, the process makes a transition from one state to another, after it has spent an amount of time – state staying time, on the state it starts from. This state staying time is exponentially distributed with some rate. When the process leaves one state, it will next enter another state with some probability independent of the time spent at the previous state. Similar as for the "cell phone adoption" example, the rate and the transition probability are essential and well-captured factors in the CTMC. In this paper, we propose a rate-based information flow model based on the CTMC framework.

4.2 Rate-based Information Flow Model

In this subsection, we propose a rate-based information flow model on a network $G(n, w, \tau)$ based on the CTMC, in which each node is a state, the weight is represented as the transition probability, and the delay is represented as the staying time in each state. Figure 2 illustrates an example of our model. We assume that the information stays in a node *i* for a certain time period T_i before making a transition to others. Then information flows to other nodes *j, k,* and *l* according to transition probabilities P_{ij}, P_{ik}, and P_{il}.

In the rest of this subsection, we first describe how to estimate the staying time in each state as well as the transition probability. Then we summarize the proposed rate-based information flow model. Afterwards, we propose a recommendation algorithm and a ranking algorithm based on the proposed rate-based information flow model.

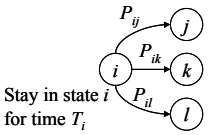


Figure 2: Rate-based Information Flow Model

4.2.1 Out-State Rate Estimation

We assume that the staying time at node *i* follows an exponential distribution with out-state rate q_i. According to the property of the exponential distribution, the expected value of an exponentially distributed random variable X_i with rate q_i is given by

$$E(X_i) = \frac{1}{q_i} = T_i \quad (10)$$

Therefore, we estimate the out-state rate by the expected value of the observations of the staying time at each node in the network.

4.2.2 Transition Probability

We estimate the transition probability based on the instances on the inter-state transition and the time delay on each transition. Given the out-state rate, we estimate the transition probability from user *i* to user *j* as

$$P_{ij} = \sum_c q_i \exp\left(-q_i t_{ij(c)}\right) \quad (11)$$

where $t_{ij(c)}$ is defined as the inter-state diffusion time from node *i* to node *j* on *c* instances.

According to Equation (9), we have

$$q_{ij} = q_i P_{ij} \qquad (i \neq j) \quad (12)$$

which we define as the inter-personal diffusion rate from user *i* to user *j*. Thus we have all the elements in the **Q** matrix ready for use.

4.2.3 Our Proposed Information Flow Model

To summarize, the rate-based information flow model is described in Figure 3, where we estimate the out-state rate, transition probability and inter-personal diffusion rate to generate the transition rate matrix **Q**.

Algorithm Rate-based Information Flow Model

Input: $G(n, w, \tau)$: user adoption data with timestamp

Output: Q: transition rate matrix

Begin

1) Estimate the out-state rate by Equation (10)
2) Estimate transition probability by Equation (11)
3) Estimate inter-personal diffusion rate q_{ij} by (12)
4) Generate transition rate matrix **Q**

End

Figure 3: Rate-based Information Flow Model

4.3 Recommendation Algorithm

Assume we have user adoption data on *N* users and *M* items. We model user adoption behaviors by our proposed rate-based

information flow model in which information flows from some users (social leaders) to others (followers) in different rates.

Specifically, given the detailed log data, including the timestamp of each adoption, for each pair of users, we calculate the out-state rate and transition probability by comparing their adoption timestamps on the same items. Following the steps listed in Figure 3, we generate a rate-based information flow model for the user adoption data.

For the recommendation problem, given at time *t=0*, user *i* adopts an item, then the information starts to flow from this user to others in the network. We predict users' preferences of information by estimating who will most likely adopt the item by time $t = \tau$, in other words, information will flow to them. To predict users' preferences by time $t = \tau$, we estimate the probability that the information flows from user *i* to others as the probability that transition $i \to j \quad (j \neq i)$ is enabled in $[0, \tau]$ as $L(j \mid i, \tau)$, which is the $(i, j)th$ element in $L(\tau)$ with

$$L(\tau) = \int_0^\tau P(t)\,dt \tag{13}$$

where $P(t)$ is the transition probability matrix with $(i, j)th$ entry $P_{ij}(t)$.

Formally, when the state space is finite, we can estimate the transition probability by solving

$$\begin{cases} P'(t) = P(t)\mathbf{Q} \\ P(0) = \mathbf{I} \end{cases} \tag{14}$$

where $\mathbf{I}$ is the identity matrix. The solution is

$$P(t) = e^{t\mathbf{Q}} = \sum_{m=0}^{\infty} \frac{(t\mathbf{Q})^m}{m!} \tag{15}$$

If $\mathbf{Q}$ can be diagonalized by $\mathbf{Q} = \mathbf{M}\mathbf{D}\mathbf{M}^{-1}$, then

$$P(t) = \sum_{m=0}^{\infty} \frac{\mathbf{M}(t\mathbf{D})^m \mathbf{M}^{-1}}{m!} = \mathbf{M}e^{t\mathbf{D}}\mathbf{M}^{-1} \tag{16}$$

For large $\mathbf{Q}$, Taylor approximation can also be used

$$P(t) \approx \lim_{m \to \infty} \left(\mathbf{I} + \mathbf{Q}\frac{t}{m} \right)^m \tag{17}$$

Specifically, the steady state distribution of the CTMC can be calculated in the ways mentioned in the following theorem.

Theorem 1. Given an irreducible CTMC, suppose $\exists \pi_i$'*s* s.t. $\pi_i > 0$ satisfies

$$\begin{cases} \sum_i \pi_i = 1 \\ \pi\mathbf{Q} = 0 \end{cases} \tag{18}$$

π_i'*s* are then the steady state distribution for the CTMC and the CTMC is ergodic. The solution is

$$\pi = \lim_{t \to \infty} e^{t\mathbf{Q}} = \lim_{t \to \infty} \sum_{m=0}^{\infty} \frac{(t\mathbf{Q})^m}{m!} = \lim_{t \to \infty} \mathbf{M}e^{t\mathbf{D}}\mathbf{M}^{-1} \tag{19}$$

The detailed proof of this theorem can be found in [28].

To summarize, the rate-based information flow for recommendation algorithm is described in Figure 4, where we estimate how likely the transition will enable during a time period by the transition probability of the CTMC.

Algorithm Rate-based Information Flow for Recommendation

Input: **Q:** transition rate matrix
u: initial users who have adopted the recent item at time *0*
n: all the users, including sets $\mathbf{u}$ and $\mathbf{z} = \mathbf{n} / \mathbf{u}$
τ : when the recommendation will be made to users

Output: $L(\mathbf{z} \mid \mathbf{u}, \tau)$ **:** how likely other users will adopt the item by time τ

Begin

1) Matrix diagonalization: $\mathbf{Q} = \mathbf{M}\mathbf{D}\mathbf{M}^{-1}$
2) Estimate $P(\tau)$ by Equation (16) or (17)
3) Estimate $L(\tau)$ by Equation (13)
4) Given a group of user *u*, estimate $L(\mathbf{z} \mid \mathbf{u}, \tau)$ by

$$L(\mathbf{z} \mid \mathbf{u}, \tau) = \sum_{u_k \in u} L(\mathbf{z} \mid u_i, \tau) \tag{20}$$

End

Figure 4: Rate-based Information Flow for Recommendation Algorithm

4.4 Ranking Algorithm

Similarly we assume we are given the user adoption data as described in Section 4.3. For the ranking problem, we pose the problem as if an arbitrary user $j \ (j \neq i)$ adopts an item at *t=0*, when will be the average time in which user *i* adopts it? In Continuous-Time Markov Chain, this question can be answered by the mean first-passage-time.

Let $\mathbf{M}$ be the first-passage time matrix of the CTMC with the (i, j) *th* element as m_{ij}. The mean first passage time m_{ij} from *i* to *j* is defined as the expected time taken until the first arrival at node *j* starting at node *i*.

Let *v* be any constant such that $v \geq \max_i(q_i)$. Divide the off-diagonal components of $\mathbf{Q}$ ($q_i \mathrm{P}_{ij} \quad (i \neq j)$) by *v* and replace its diagonal components $-q_i$ by $1 - q_i / v$, we then have a uniformized chain (discrete time), whose transition matrix $\mathbf{P}_v$ can be related to $\mathbf{Q}$ though *v* as follows

$$\mathbf{P}_v = \mathbf{I} + \frac{1}{v}\mathbf{Q} \tag{21}$$

Let $\mathbf{M}_\lambda$ denote the matrix of the first passage time of the uniformized chain (discrete time). According to [26], the mean first-passage time matrix of discrete-time Markov chain is given by

$$\mathbf{M}_v = \left(\mathbf{I} - \mathbf{Z}_v + \mathbf{E}(\mathbf{Z}_v)_{dg} \right)\mathbf{D} \tag{22}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{E}$ is a matrix containing all ones, and $\mathbf{D}$ is the diagonal matrix with elements $d_{ii} = \frac{1}{\pi(i)}$ where $\pi(i)$ is the steady state distribution of node *I* in this discrete-time Markov chain, $\mathbf{Z}_v$ is the fundamental matrix [30] of this discrete-

time Markov chain with

$$\mathbf{Z}_v = \left(\mathbf{I} - \mathbf{P}_v + \mathbf{P}_v^{\infty}\right)^{-1} \quad (23)$$

$\left(\mathbf{Z}_v\right)_{dg}$ results from $\mathbf{Z}_v$ by setting off-diagonal entries to zero, and $\mathbf{P}_v^{\infty}$ is the limiting matrix of $\mathbf{P}_v$ with each row of $\mathbf{P}_v^{\infty}$ as π^T, or $\mathbf{P}_v^{\infty} = \mathbf{e}\boldsymbol{\pi}^T$, where $\mathbf{e}$ is a column vector with all ones. The matrix of the first-passage time of the original Continuous-Time Markov Chain $\mathbf{M}$ is

$$\mathbf{M} = \frac{1}{v}\left(\mathbf{M}_v\right)_{of} + \boldsymbol{\Lambda}\left(\mathbf{M}_v\right)_{dg} \quad (24)$$

where $\boldsymbol{\Lambda} = diag\left(q_i^{-1}\right)$, $\left(\mathbf{M}_v\right)_{dg}$ results from $\mathbf{M}_v$ by setting off-diagonal entries to zero, and $\left(\mathbf{M}_v\right)_{of}$ results from $\mathbf{M}_v$ by setting diagonal entries to zero.

The computational complexity is as high as $O\left(|N|^3\right)$ because we need to invert a matrix of size $|N| \times |N|$ in Equation (23). Fortunately, an efficient way of calculating first passage time in a large Continuous-Time Markov Chain by means of Laplacian transforms is discussed in [27]. Since the computational efficiency is out of the scope of this paper, we refer to [27] for more details.

Thus, the rank score for each user j is estimated as

$$R(j) = \frac{1}{\frac{1}{|N-1|}\sum_{i \neq j} m_{ij}} \quad (25)$$

To summarize, the rate-based information flow for ranking algorithm which we call the DiffusionRank Algorithm is described in Figure 5, where the mean first-passage time of the CTMC is utilized to rank the users by the flow efficiency.

Algorithm DiffusionRank Algorithm

Input: **Q:** transition rate matrix
n: users

Output: $R(\mathbf{n})$ **:** rank scores

Begin

1) Generate a uniformized matrix with transition matrix as $\mathbf{P}_v$ according to Equation (21)
2) Generate the steady state distribution of $\mathbf{P}_v$
3) Calculate the mean first passage matrix $\mathbf{M}_v$ of the uniformized chain according to Equation (22)
4) Calculate the mean first passage time matrix $\mathbf{M}$ of the original CTMC according to Equation (24)
5) Calculate the rank scores $R(\mathbf{n})$ by Equation (25)

End

Figure 5: DiffusionRank Algorithm

5. EXPERIMENTS

In this section, we first describe the datasets used to evaluate our algorithms. We observe how information flow model captures user adoption patterns. Following that, we analyze the diffusion time from global and structural perspectives. Afterwards, we demonstrate the experimental results on both recommendation and ranking.

5.1 Experiment Set-Up

We demonstrate our proposed algorithms on two datasets. The first dataset is collected from NEC's "EigyoRyoku 21" (denoted as ER and stands for Sales-Force in Japanese) system. The ER system is a knowledgebase to support sales staffs with registered documents that include articles, slides, *etc*. We collected a thirty-month period of clickstream log files from April 1 2004 to September 19 2006 covering 3,528 users and 31,379 documents. Nine user actions are identified: {"Login", "Register_Feedback", "Preview", "Abstract", "Document Download", "Search", "Register", "Update", "Delete"}. The clickstream log is partitioned into sessions that start with "login" followed by a sequence of user actions. The timestamps of users' actions as well as the disclosure of the documents are included in the dataset. In this paper, we assume if a user accesses a document (including "Preview", "Abstract" or "Download"), he or she has adopted it.

The second dataset is MovieLens dataset [4], which consists of 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000. Each user has rated at least 20 movies. We assume the user has adopted the movie as long as this user ever provided a rate on the movie.

5.2 User Adoption Patterns

Based on ER dataset, we simulate the user adoption process to illustrate user adoption patterns. Figure 6 illustrates the adoption process on one document. We first cluster users into five groups according to Rogers' diffusion of innovation theory [1] (the user clusters from the left to the right are innovators, early adopters, early majority, late majority and laggards respectively) according to how early their adoptions are in the dataset. These user groups are illustrated as half-rings in the figure. Each node represents a user, with user ID as the label. When a document is disclosed and users start to adopt it, we mark the nodes of these users as red shaded boxes accordingly based on the user's relative adoption time. We do observe consistent sequential patterns from user adoption data, which confirm that information flows from innovators to early adopters, early majority, late majority, until laggards.

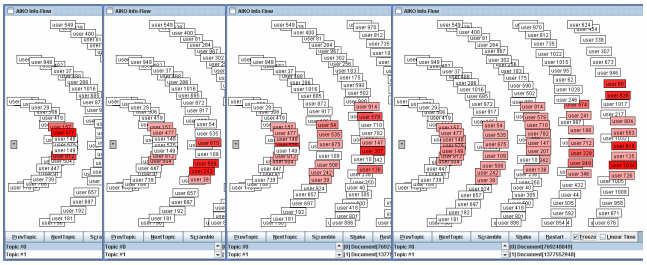


Figure 6: Visualization on user adoption patterns. This series captures the fact that after one document being disclosed (the small rectangular on the left side), how users adopt it over time. Each node represents a user, with user ID as the label. The nodes are marked by red shaded boxes when the corresponding users adopt the document

5.3 Diffusion Time: Case Study

To analyze how the information flows in different rates in these two datasets, we study the distributions of global diffusion time, and structural diffusion time.

5.3.1 *Global Diffusion Time*

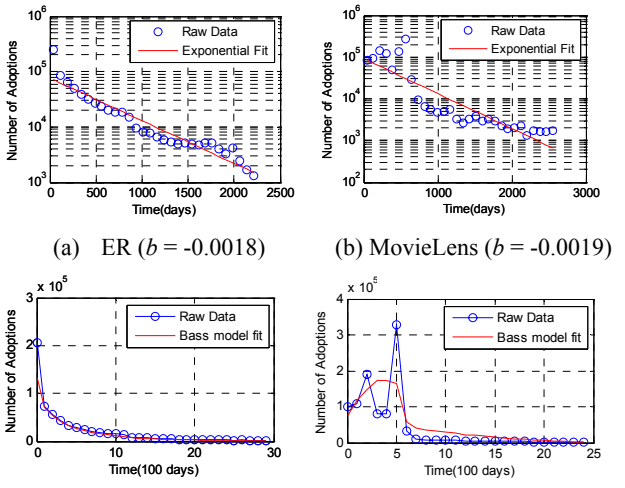
We define the global diffusion time as how long it takes that information spreads through the population. This concept is in the inverse proportion of the global diffusion rate which has been captured in diffusion of innovation theory [6] and Bass model [15].

For both datasets, we ranked the user adoption behaviors by the adoption time. The global diffusion time is calculated as the difference of the disclosure time and the adoption time, keeping the relative position unchanged. Intuitively, we expect there are a larger amount of adoptions at the beginning and followed with decay because the value of the items decays over time. Figures 7(a) and (b) illustrate how the number of adoptions of the items changes over time. It does confirm our expectation. A large amount of adoptions take place at the early stage when the items are just disclosed, rapidly decaying afterwards for both datasets. To illustrate the decay rate, we fit the data to an exponential function,

$$Y = a\exp(b \cdot X) \tag{26}$$

As indicated by the coefficient of exponential fit, the decay rate in the ER dataset (b = -0.0018) is similar as that in the MovieLens dataset (b = -0.0019).

To further analyze how the information diffuses over time, we utilize the Bass model [15], which quantifies the adoption of innovations by estimating the introduction and acceptance rate variables. Figures 7 (c) and (d) illustrate how the Bass model fits two datasets by using Equation (1). Comparing the innovation coefficients of these two cases, (p = 0.2748 for ER and p = 0.0756 for MovieLens), we can tell that ER system has higher percentage of innovators than MovieLens system, which is reasonable because the documents in ER system are more time-sensitive than the movies in MovieLens system.



(a) ER (b = -0.0018) (b) MovieLens (b = -0.0019)

(c) ER (p = 0.2748) (d) MovieLens (p = 0.0756)

Figure 7: The Distributions of global diffusion time on ER and MovieLens datasets

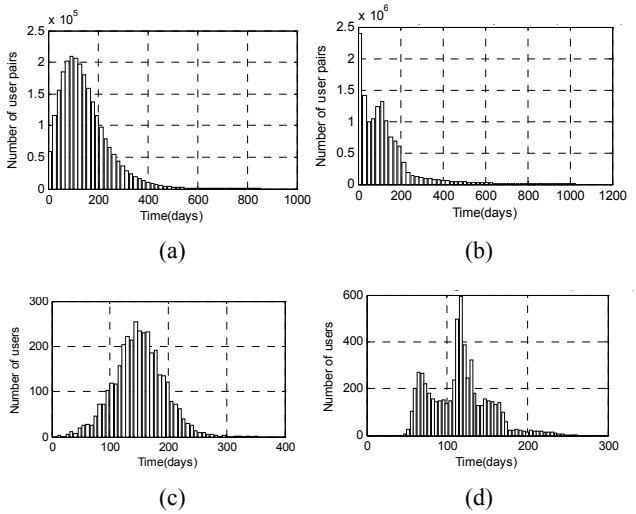
Note: For Figures 7-9, the x-axis represents time (days or 100 days), y-axis represents the number of corresponding diffusion times located in a certain time.

The global diffusion time provides us with an overall view of how quickly the information diffuses for different products or technologies, thus offers a means to look further into the special insights and properties of the products or technologies. However, we still have no clue on how the information flows to different nodes in the network without looking into the structural level.

5.3.2 *Structural Diffusion Time*

We define the structural diffusion times as those diffusion times which take the network structure into account. The structural diffusion times include inter-personal diffusion time, which we define as how long it takes that information spreads from one node to another, and state staying time, which is defined in Section 4.1. In this paper, we make the same assumption as we did in [5]: the fact that two users access the same item sequentially is modeled as an information flow process: the information is flowing from early adopters to late adopters.

For both datasets, we calculate the difference of adoption time of each pair of users on the same item as the inter-personal diffusion time. For each pair of users, we take the average over the common items both users adopt to indicate how rapidly the information flows from one user to another on average. Figures 8 (a) and (b) illustrate the distributions of the average inter-personal diffusion time in ER and MovieLens datasets. For both datasets, the average inter-personal diffusion times vary from several days to several hundred days. We also study the distribution of state staying time at each user. For each user, we calculate the adoption time difference of him/her and others and take the average to indicate how rapidly the information flows out from this node to others. Figures 8 (c) and (d) illustrate the distributions of the average state staying time in ER and MovieLens datasets. Again, for both datasets, the average state-staying times vary from several days to several hundred days. The variety reminds us of the example of "cell phone adoption". In real situations, some users more efficiently influence other users than some others do. Users with similar properties as the "father" and "grandmother" do exist. How to differentiate them will affect various applications such as recommendation and ranking.



(a) (b)

(c) (d)

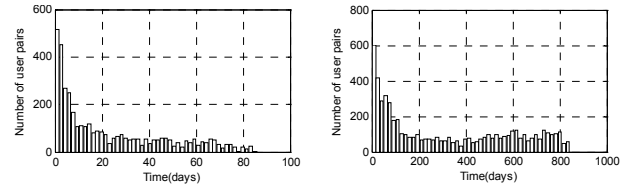
Figure 8: The distributions of the structural diffusion time

The distributions of the average inter-personal diffusion times on ER dataset (a) and MovieLens dataset (b).

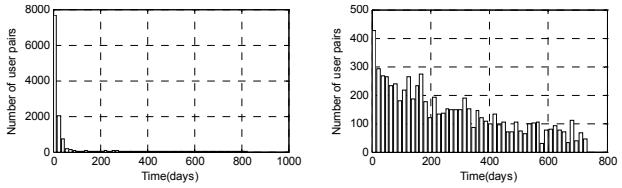
The distributions of the average state staying time on ER dataset (c) and MovieLens dataset (d)

To look closer, Figure 9 provides the individual cases for the state staying time. Figures 9(a) and (b) compare the state staying time for user 3526 and user 2433 on ER dataset. The state staying time for user 3526 ranges from one day to around 80 days; while

for user 2433, it ranges from one day to more than 800 days. Thus, on average, user 3526 influences others much more efficiently than user 2433. The similar conclusion can be drawn by comparing the state staying time of user 2292 and user 2980 on MovieLens dataset as illustrated in Figures 9(c) and (d). Also we can see that modeling of the state staying time as exponential distribution is reasonable, although not perfect.



(a) ER: UserID 3526 (b) ER: UserID 2433



(c) MovieLens: UserID 2292 (d) MovieLens: UserID 2980

Figure 9: The case study of distribution of the state staying time

5.4 Recommendation Performance

To demonstrate the performance of our recommendation algorithm, for the ER dataset, we first divide both datasets into two sets: the training set and the test set. The data from April 2004 to April 2005 serve as the training data, and the data from May to July 2005 serve as the test data. To exclude casual users who had very few activities, we selected 1170 active users who adopted more than 50 documents in the training period; and more than 10 documents in the test period in our experiments. In total, there are 23,894 documents involved in this selected dataset. 201,750 adoption actions were recorded. The average number of adoption actions per user is 172, and the average number of actions per document is 8. For the MovieLens dataset, we select first 80% as the training set and later 20% as the test set.

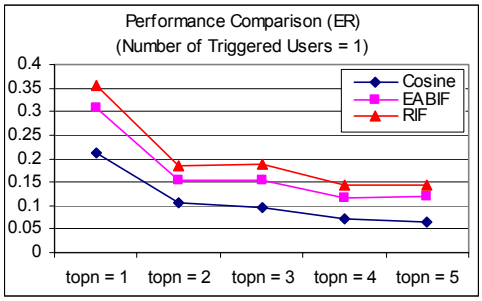
In the experiments, we simulate the situation of adoption behaviors. There are 586 documents disclosed during the test period from May to July 2005. The mean value of the number of users who adopted these 586 documents during May to July 2005 is 18. These documents were first adopted by one or multiple users – innovators. We then predict who else will most likely adopt the documents following these early adopters within 30 days. This strategy is suitable for online document/product pushing service or advertisement – recommending the items to potential customers according to the choices of innovators. It can also be used in the traditional recommendation scenario by estimating how likely one user will be interested in the documents and recommending those top-ranked items to him/her.

To evaluate the accuracy of predictions, we measure the average recommendation accuracy, which represents the percentage of them who are actually interested in the item among the recommended users.

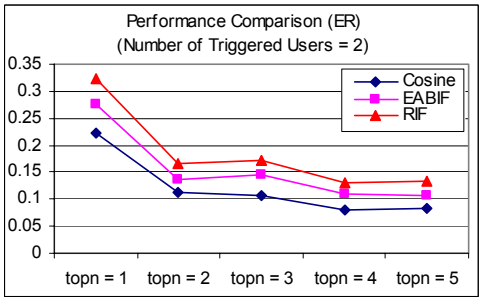
In our experiments, we compare the performance of the following algorithms:

1. *Collaborative Filtering based on Cosine Similarity (baseline) (denoted as Cosine)*
2. *Early adoption based information flow model as proposed in [5] (denoted as EABIF)*
3. *Rate-based information flow model (proposed in this paper, and denoted as RIF)*

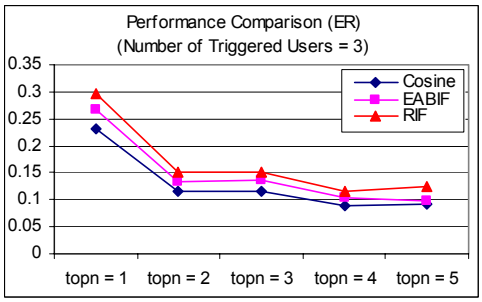
Figure 10 compares the average recommendation accuracy of Collaborative Filtering based on Cosine Similarity (labeled by CF), early adoption based information flow model (labeled by EABIF) and rate-based information flow model (labeled by RIF) on both ER ((a)–(c)) and MovieLens datasets ((d)-(e)). We demonstrate the results in the situations when one or multiple users adopted the item within 30 days. In both datasets, RIF beats CF and EABIF. In ER dataset, RIF improves accuracy by 67% comparing to CF, and 20% comparing to EABIF. In MovieLens dataset, RIF improves accuracy by 80% comparing to CF, and 53% comparing to EABIF. We also find that for MovieLens dataset, given the recommendation time as 30 day, sometimes the EABIF algorithm can not perform as well as CF does. The reason is that EABIF looks for potential users in a much longer duration – sometimes these potential users are laggards ("grandma" in our cell phone adoption example). However, the interest of these laggards on the item may have not been triggered yet during a short time period like 30 days.



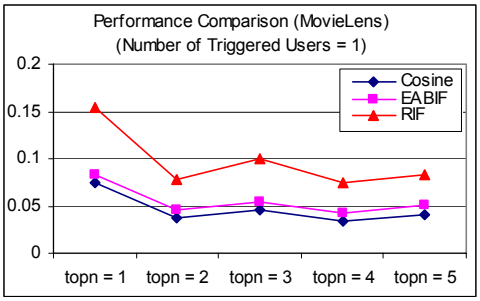
(a) ER: Number of adopted users = 1



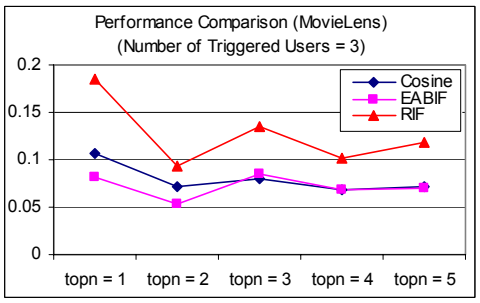
(b) ER: Number of adopted users = 2



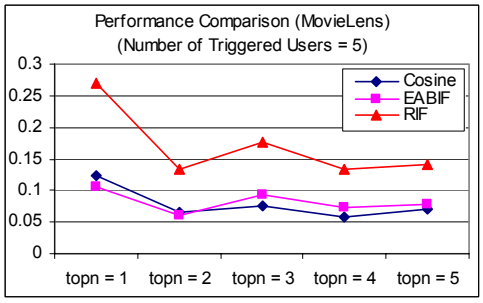
(c) ER: Number of adopted users = 3



(d) MovieLens: Number of adopted users = 1



(e) MovieLens: Number of adopted users = 3



(f) MovieLens: Number of adopted users = 5

Figure 10: Average recommendation accuracy of RIF comparing to CF and EABIF on ER and MovieLens datasets given different number of adopted users

5.5 Ranking Performance

To demonstrate the performance of our ranking algorithm, we compare the performance of the following algorithms:

1. *PageRank (with the damping factor as 0.85)*
2. *HITS*
3. *DiffusionRank (our proposed algorithm)*

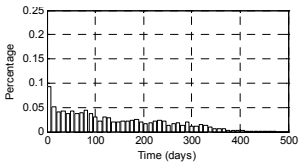
Table 1 and Table 2 show the top 10 ranked users in ER and MovieLens datasets using these algorithms. To further analyze the result, we illustrate the in-node diffusion time for 1st ranked user in PageRank, and DiffusionRank in Figure 11. We define the in-node diffusion time as how long it takes that information flows from other nodes to this particular node. The in-node diffusion time is an important indicator which is related to the flow efficiency to a particular node although not the only one. We can see that on average, the in-node diffusion time for the 1st ranked user in DiffusionRank (Figure 11(b)) is much smaller than that of the 1st ranked in PageRank and HITS (Figure 11(a)). This demonstrates that the information flows more rapidly to these users who ranked higher in DiffusionRank than to these higher ranked in PageRank and HITS. A similar conclusion can be drawn from the result on MovieLens dataset (Figures 11(c) and (d)).

Table 1. Top 10 ranked users in ER dataset

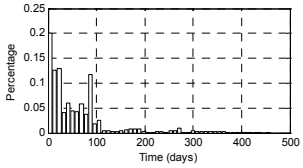
Rank	User IDs ordering			
	PageRank	HITS		DiffusionRank
		Authority	Hub	
1	31	31	7	722
2	5	5	31	563
3	9	9	1016	1
4	7	10	10	66
5	10	7	103	469
6	246	246	5	673
7	363	363	9	582
8	970	970	29	952
9	29	199	199	6
10	366	29	239	849

Table 2. Top 10 ranked users in MovieLens dataset

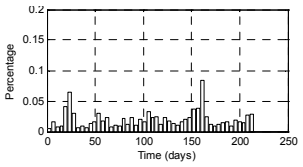
Rank	User IDs ordering			
	PageRank	HITS		DiffusionRank
		Authority	Hub	
1	683	293	276	1
2	729	94	303	189
3	416	655	92	188
4	796	234	222	291
5	551	682	804	468
6	189	7	268	221
7	56	59	749	342
8	234	796	130	796
9	532	308	194	919
10	807	551	378	56



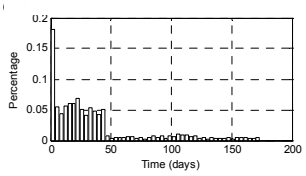
(a) ER: UserID 31



(b) ER: UserID 722



(c) MovieLens: UserID 683



(d) MovieLens: UserID 1

Figure 11: Distribution of the in-node diffusion times of top-ranked users. The x-axis represents time (days), y-axis represent the percentage of the in-node diffusion times to this user located in a certain time

6. CONCLUSIONS AND FUTURE WORK

Social influence is the process whereby people directly or indirectly influence the thoughts, feelings and actions of others. PageRank and HITS use the hyperlink structure of the web pages to determine their "importance". Word-of-mouth marketing and recommendation is another way of leveraging social influences.

However, none of these approaches take the diffusion rates in the network into account.

In this paper, we propose a novel Information Flow model that captures the diffusion rates of information in a network. By modeling the diffusion of information flow, we can tackle two refined problems in recommendation and ranking. First, we can measure how likely information will flow from a specific sender to a specific receiver within a limited time, and thus predict who will be the most likely recipients of information in a recommendation. Second, we can estimate the expected time for information to diffuse to a particular user in the network, and thus rank users based on how quickly information will travel to them.

Our recommendation and ranking algorithms use the transient transition probability and the mean first-passage time in a Continuous-Time Markov Chain. Consequently, we can address the diffusion efficiency of information flow on the inter-personal and individual levels.

In our experiments, the diffusion efficiency of the proposed algorithms is demonstrated for both the prediction of automatic recommendations and ranking of influential users. For recommendation, we chose a recommendation period of 30 days. Compared with traditional Collaborative Filtering and the uniform-diffusion information flow model (EABIF), our rate-based information flow (RIF) algorithm improves 67% and 20% respectively for recommendation accuracy in the ER dataset. In the MovieLens dataset, RIF improves accuracy by 80% comparing to CF, and 53% comparing to EABIF. Furthermore compared with PageRank and HITS algorithms, our proposed DiffusionRank ranks users based on how efficiently information will flow to them in the network.

Ongoing work includes formally defining "influence" between users and conducting our proposed algorithms on datasets with explicit links, such as blog datasets, to track the influence flow. Also, some recent evidences demonstrate that heavy tailed statistics happen in some human actions [22-25]. Thus, another direction is to estimate the state staying time by hyper-exponential distribution to fit the heavy tailed distribution. Then we can build the rate-based information flow model based on semi-Markov process [29].

7. ACKNOWLEDGEMENTS

We would like to thank Ming-Ting Sun, Ching-Yung Lin, Dengyong Zhou, Hari Sundaram, and John Lafferty for valuable discussions and comments at the early stage of this paper. We also thank Yves Petinot for help on revising the paper.

8. REFERENCES

[1] R. B. Cialdini, Influence: Science and Practice, Apr 2003.

[2] U. Shardanand, and P. Maes, Social Information Filtering: Algorithms for Automating "Word of Mouth", CHI '95.

[3] G. Adomavicius, and A. Tuzhilin, Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, IEEE Trans. Knowl. Data Eng. 17(6): 734-749, 2005.

[4] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, Grouplens: An open architecture for collaborative filtering of netnews. In Proc. of the ACM Conference on Computer Supported Cooperative Work: 175-186, 1994.

[5] X. Song, B. L. Tseng, C.-Y. Lin, and M.-T. Sun, Personalized Recommendation Driven by Information Flow, International ACM SIGIR Conference on Research & Development on Information Retrieval, Seattle, August 6-11, 2006.

[6] E. M. Rogers, Diffusion of Innovations, The Free Press: New York, 1995.

[7] V. Mahajan, E. Muller, F. Bass. New Product Diffusion Models in Marketing: A Review and Directions for Research. Journal of Marketing 54:1, pp. 1-26, 1990.

[8] A. Pucci and M. Gori, Random-Walk Based Scoring Algorithm with Application to Recommender Systems for Large-Scale E-Commerce, WEBKDD 2006.

[9] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. Computer Networks, 30(1-7):107-117, 1998.

[10] A. N. Langville and C. D. Meyer. Deeper inside PageRank. Internet Mathematics, 1(3):335-400, 2004.

[11] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604-632, 1999.

[12] K. Berberich, M. Vazirgiannis, and G. Weikum, *T-Rank: Time-aware Authority Ranking*, 3rd Workshop on Algorithms and Models for the Web-Graph, October 2004.

[13] S. White and P. Smyth. Algorithms for estimating relative importance in networks. In Proc. of ACM SIGKDD Conf., 2003.

[14] E. Adar, L. Zhang, L. A. Adamic, R. M. Lukose, Implicit Structure and the Dynamics of Blogspace, Workshop on the Weblogging Ecosystem, May 18th, 2004.

[15] F. Bass, A new product growth for model consumer durables, *Management Science* 15 (5): p215-227, 1969.

[16] S. Hill, F. Provost, and C. Volinsky, Network-Based Marketing: Identifying Likely Adopters via Consumer Networks, Statist. Sci. 21, no. 2, 256–276, 2006.

[17] M. Richardson and P. Domingos, "Mining Knowledge-Sharing Sites for Viral Marketing," KDD 2002.

[18] P. Domingos and M. Richardson, "Mining the Network Value of Customers," KDD 2001.

[19] D. Kempe, J. Kleinberg, E. Tardos. Maximizing the Spread of Influence through a Social Network. Proc. 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2003.

[20] D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins**,** Information diffusion through blogspace, WWW 2004.

[21] R. E. Kraut, R. E. Rice, C. Cool, and R. S. Fish, Varieties of Social Influence: The Role of Utility and Norms in the Success of a New Communication Medium, *Organization Science*, Vol. 9, No. 4, pp. 437-453, Jul. - Aug 1998.

[22] Z. Dezso, E. Almaas, A. Lukacs, B. Racz, I. Szakadat, A.-L. Barabasi, *Dynamics of information access on the web* Physical Review E 73 (6): Art. No. 066132 Part 2, 2006.

[23] C. Dewes, A. Wichmann, and A. Feldmann, *An analysis of Internet chat systems.* ACM/SIGCOMM Internet Measurement Conference, 2003.

[24] *J. G. Oliveira and A.-L. Barabási,* Human dynamics: Darwin and Einstein correspondence patterns, *Nature 437, 1251 2005.*

[25] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor & A.-L. Barabási, Modeling bursts and heavy tails in human dynamics Phys. Rev. E 73, 036127, 2006.

[26] D. D. Yao, First-Passage time Moments of Markov processes, *Journal of Applied Probability*, Vol. 22, No. 4, pp. 939-945, Dec., 1985.

[27] P. Harrison, W. Knottenbelt, Passage Time Distributions in Large Markov Chains, ACM International Conference on Measurement and Modeling of Computer Systems, pp.77-85, June, 2002.

[28] J. R. Norris, Markov Chains, Cambridge University Press, 1997.

[29] S. M. Ross. Introduction to probability models. Academic Press, New York, 2003.

[30] I. G. Kemeny and J. L. Snell, Finite Markov Chains, Springer-Verlag. New York, 1976.