Information Sciences 278 (2014) 535–544

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Maximizing the spread of influence ranking in social networks

Tian Zhu [a,b,*], Bai Wang [b], Bin Wu [b], Chuanxi Zhu [c]

[a] National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, PR China
[b] Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, PR China
[c] School of Science, Nanchang University, Nanchang, Jiangxi 330031, PR China

ARTICLE INFO

*Article history:*
Received 1 March 2010
Received in revised form 7 November 2013
Accepted 12 March 2014
Available online 25 March 2014

*Keywords:*
Data mining
Social network
Influence maximization
Information propagation
Node centrality

ABSTRACT

Information flows in a network where individuals influence each other. In this paper, we study the influence maximization problem of finding a small subset of nodes in a social network that could maximize the spread of influence. We propose a novel information diffusion model *CTMC-ICM*, which introduces the theory of Continuous-Time Markov Chain (CTMC) into the Independent Cascade Model (ICM). Furthermore, we propose a new ranking metric named SpreadRank generalized by the new information propagation model CTMC-ICM. We experimentally demonstrate the new ranking method that can, in general, extract nontrivial nodes as an influential node set that maximizes the spread of information in a social network and is more efficient than a distance-based centrality.



## 1. Introduction

A social network – the graph of the relationships and interactions within a group of individuals – plays a fundamental role as a medium for the spread of information, ideas, and influence among its members [10,22,2]. Social network analysis has received increasing interest in many different areas in recent years, including community detection [14,5,8,3], role detection [7,17], etc. In this paper, we present our work toward addressing one of the challenges in social network analysis, namely efficiently finding influential individuals in a social network [15,20].

We consider the following hypothetical scenario as a motivating example. A telecommunication operator develops a new potential product and wants to market it through their customer network. The company selects and expects a small number of initial users in the network to use it for free. The company's goal is to have these initial users appreciate the product and start influencing their friends and the friends of their friends. Thus, a large population in the customer network would adopt the application through the word-of-mouth effect. The problem is determining who to select as the initial users so that they eventually influence the widest variety and the largest number of people in the network, i.e., the problem of determining the influential individuals in a social network.

This problem is referred to as influence maximization, the solving of which would be of interest to many companies besides the telecommunication operators that are promoting their products, services, and innovative ideas through the powerful word-of-mouth effect (also known as viral marketing) [10].

To investigate this problem, we must use a model of information diffusion in a social network. There are two basic diffusion models: the Linear Threshold Model (LTM) and the Independent Cascade Model (ICM) [10]. Both LTM and ICM are stochastic models in which information flows from a node to its neighboring nodes at each time-step according to some

* Corresponding author at: No. 3 Jia Yumin Road, Chaoyang District, Beijing 100029, PR China. Tel.: +86 13811878704; fax: +86 10 82990399.
*E-mail address:* zhutian@cert.org.cn (T. Zhu).

http://dx.doi.org/10.1016/j.ins.2014.03.070
0020-0255/

probabilistic rule. Therefore, when considering the problem of finding sets of influential nodes in a social network based on LTM or ICM, we must compute the expected number $\sigma(A)$ of nodes influenced by a given set A of nodes. Determining an efficient method to compute $\sigma(A)$ exactly remains an open question, and good estimates were obtained by simulating the random process many times [10].

In this paper, we propose a novel information diffusion model derived from the basic ICM that efficiently computes a good estimate of $\sigma(A)$. Then, we propose a new ranking metric named SpreadRank provided by the new information propagation model. We experimentally demonstrated that the new model can provide good approximations for finding sets of influence nodes in a social network. We also demonstrated that the new ranking method can, in general, extract the nontrivial nodes as influential nodes more accurately than degree centrality and more efficiently than distance-based centrality.

The remainder of the paper is organized as follows. We briefly review the related work in Section 2. Section 3 describes the new information model adaptable to both undirected and directed networks. In Section 4, we propose the new ranking metric SpreadRank derived from the new information propagation model. In Section 5, we describe the experimental results. Finally, conclusions and future work are given in Section 6.

## 2. Related work

Domingos and Richarson [4,16] first posed influence maximization as a fundamental algorithmic problem. They considered the question as a probabilistic model of interaction. Then, Kempe et al. [10] formulated the problem as a discrete optimization problem and presented an extensive study of this problem for the first time. They demonstrated that approximation algorithms for maximizing the spread of influence in optimization models can be developed in a general framework based on submodular functions. They also provided computational experiments on large collaboration networks, demonstrating that their algorithms significantly out-performed node-selection heuristics based on the well-studied notions of degree centrality and distance centrality from the field of social network analysis.

There are also many recent studies aimed at addressing this interesting problem. Java et al. [9] used the basic Linear Threshold Model proposed by Kempe et al. to select an influential set of bloggers to maximize the spread of information on the blogosphere. Kimura and Saito [11]proposed the shortest-path based on the influence cascade model and provided efficient algorithms for computing the spread of influence under this model.

The two basic models are described as follows. In the basic Linear Threshold Model, start with an initial set of active nodes $A_0$, with each node having a certain threshold for adopting an idea of being influenced. The node becomes activated if the sum of the weights of the active neighbors exceeds the threshold. Thus, if node $v$ has a threshold $\theta_v$ and an edge weight $b_{wv}$ such that neighbor w influences $v$, then $v$ becomes active only if

$$\sum_{w\ active\ neighbors\ of\ v} b_{wv} \geqslant \theta_v. \tag{1}$$

In the basic Independent Cascade Model, again begin with an initial set of active nodes $A_0$, and the process unfolds in discrete steps according to the following randomized rule. When node $v$ first becomes active in step $t$, it is given a single chance to activate each currently inactive neighbor $w$; it succeeds with a probability $p_{v,w}$ – a parameter of the system – independently of the history thus far. (If $w$ has multiple newly activated neighbors, their attempts are sequenced in an arbitrary order.) If $v$ succeeds, then $w$ will become active in step $t+1$; but whether $v$ succeeds or not, it cannot make any further attempts to activate $w$ in subsequent rounds. Again, the process runs until no additional activations are possible.

We call a node active if it has accepted the information. We assume that nodes can switch from being inactive to being active but cannot switch from being active to being inactive.

## 3. Improving the greedy algorithm for the Independent Cascade Model

In this section, we discuss our improvement of the greedy algorithm proposed by Kempe et al. [10] for the Independent Cascade Model.

### 3.1. Continuous-Time Markov Chain

**Definition 1.** A Continuous-Time Markov Chain (CTMC) [18] is a continuous time stochastic process $X(t), t \geqslant 0$ s.t. $\forall\ s,\ t \geqslant 0$, and $\forall\ i,\ j,\ x(h)$

$$P\{X(t+s)=j|X(t)=i, X(h)=x(h), 0\leqslant h\leqslant t\} = P\{X(t+s)=j|X(t)=i\}. \tag{2}$$

A Continuous-Time Markov Chain satisfies the Markov property and takes values 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 entire history of the process up to and including time $t$, depends only on the state transition probabilities, which are independent from the initial time $t$, i.e., the chain is time-homogeneous, and we denote $P_{ij}(S)$ as the transition probability from $i$ to $j$ over s time period.