The Roles of Social Network Mavens

Hussah Albinali*, Meng Han*, Jinhao Wang†, Hong Gao‡ and Yingshu Li*

*Department of Computer Science, Georgia State University, Atlanta, Georgia, USA 30303
Email: halbinalii@student.gsu.edu, mhan@cs.gsu.edu, yili@gsu.edu
†High Performance Computing Center, The Academy of Fundamental and Interdisciplinary Sciences
Harbin, China 150080, Email: wangjinbao@hit.edu.cn
‡School of Computer Science and Technology, Harbin Institute of Technology
Harbin, China 150080, Email: honggao@hit.edu.cn

Abstract—This paper studies social influence from the perspective of users’ characteristics. The importance of users’ characteristics in word-of-mouth applications has been emphasized in economics and marketing fields. We model a category of users called mavens where their unique characteristics nominate them to be the preferable seeds in viral marketing applications. In addition, we developed and verified methods to learn their characteristics from a real dataset. Also, we illustrated ways to maximize information flow through mavens in social networks. Our experiments show that our model successfully detected mavens as well as fulfilled significant roles in maximizing the information flow in a social network where Mavens considerably outperformed the spread that resulted of traditional influencer users called mavens where their unique characteristics nominated them to be the preferable seeds in viral marketing applications. These results showed the compatibility of our model with real marketing approaches.

I. INTRODUCTION

The influence maximization problem has recently generated much research. Essentially, this problem attempts to identify a small subset of nodes or seeds in a given social network, and it is expected that these seeds will maximize the spread of ideas, products, or messages, cause them to go viral. In other words, the influence maximization problem seeks a subset with size k of seed users who have the ability to maximize the spread of influence among other users in social networks. In fact, viral marketing is one of the most important applications of influence maximization problem [24],[23]. A viral marketing technique is essentially based on a network model where users exchange their knowledge and experiences throughout preexisting social relationships [3]. Kempe Goyal et al. [23] first formulated influence maximization as an optimization problem in addition to modeling the diffusion of the influence in two influence cascade models: “the linear threshold (LT) model and the independent cascade (IC) model”. They also proved that the influence maximization problem under both Linear Threshold Model and the Independent Cascade model are NP-hard. This further explain the hardness of influence maximization in social network especially in large scale social network [16]. Moreover, Goyal et al. introduced the first work to define users’ influence in a social network and developed an influence models based on users’ actions log. However, analyses of users action log in influence estimation have not been involved in most of the influence maximization research. The reason behind this elimination is that the main focus has been on the relationship [?] between users in social networks rather than users’ behavior itself. In fact, users’ properties have been overlooked in the majority of influence maximization research even though users’ activities are the foundation in the diffusion process. On the other hand, few works have been developed based on users’ activity log on the social network. For instance, Barbieris contribution in [2] is worthwhile to be mentioned as they introduced a unique influence propagation model, AIR, which is based on users’ properties instead of considering the ordinary user-to-user influence. In particular, their model focuses on extracting users’ authoritativeness and interests in a specific topic. By highlighting only these features of users, the number of parameters of their influence model was effectively reduced. Another remarkable work is Budak et al.’s paper on classifying blog users into multiple categories including connectors, mavens, salesmen and translators [8]. In fact, those classes are very close to high potential consumers in viral marketing from marketing perspective which involves market mavens, social hubs, and salespeople [14]. However, Budak et al.’s model mainly focused on blogs, and it is not extendable to social networks such as Twitter and Facebook. Moreover, many researchers have developed many methods based on specific topics and how to utilize these features to enhance the performance of influence maximization. As mentioned above, in [2] Barbieri et al. first proposed a model form topic-aware perspective. Nevertheless, to the best of our knowledge, no effort has been invested in considering the ability of given prospective seed users to initialize the propagation process in relation to diverse aspects of topics. For example, in our daily life, we may meet many people who have a high influence on their peers but their personality does not prompt them to take the initiative. Thus, they naturally resist trying new things in different areas unless they are recommended to them by people they trust. They might also take the initiative in their field e.g., music, sport, or technology, but generally, they refuse to be the first people who are exposed to new trends. In fact, the time and effort spent to convince them to adopt a new idea or product will be extensive and probably will not be profitable. As a result of their personal characteristics, they are not the best candidates to initialize the adoption of new ideas or products, even though they have a high influence on their community [6]. On the other hand, a group of users called mavens enjoy generalizing marketplace information and have a strong motivation in disseminating their experience to others.
As a result of their personality, they have a high potential to participate in distributing product information with a high credibility compared to many other users. In the marketing field, mavens are defined as "Individuals who have information about many kinds of products, places to shop, and other facets of markets, and initiate discussions with consumers and respond to requests from consumers for market information" [13]. This group was first introduced by Feick and Price [13]. The existence of market mavens later received widespread attention in physical as well as web-based channels [4]. Nevertheless, a notable absence in influence maximization research is an emphasize on market mavens based on their activities and behavior in social networks including Myspace, Facebook, Twitter and LinkedIn. In general, the influence maximization problem targets a more robust selection of seed nodes, whereas activating them will introduce a sufficient cascade for their adoption. Moreover, each node also requires an activation cost, related to the effort and budget that will be spent in order to convince those users to adopt the message. For instance, offering a free sample of a product to non-maven users may not be attractive to them, and they could simply discard it, which is a financial loss to the advertising company. In contrast, this behavior is the opposite for mavens. Mavens usually seek these kinds of offers, and they are always interested in try new things. Thus ultimately, in seed’s selection, we care about the limited budget and we attempt to find the best and lowest possible losses of seeds selection. Comparing this novel idea of detecting mavens in social networks and selecting the seed nodes among the most influencers in a given social network to seed the propagation can be clarified as followings. The power of adopting a small number of influential people could potentially prove successful and lead to wide diffusion, but if it fails, it will lead to undesirable losses of money and resources without improvement in performance. In contrast, reaching more mavens with the same budget and encouraging them to share, will increase the likelihood of promoting a viral chain. This can improve the performance by reaching more people. Fortunately, it is desirable to detect mavens and their desire to share and to avoid the most serious pitfalls of failing to ignite the influence seed. This paper makes the following contributions:

- We first introduce an important concept the maven, which is a vital concept to represent a group of users who enjoy spanning multiple product categories and have a strong desire to broadcast their experiences with these new products.
- We redefine mavens in the social networks based on their activities on the social media, and give a theoretical analysis. This definition leads to better accuracy in predicting a specific item, which reflects real-world cascade.
- Based on detecting mavens in a social network, we introduce a network model and develop a heuristic method to maximize the information diffusion in a social network by maximizing the multi-commodity flow of information in the network based on the proposed model.

- Last but not the least, we tested and verified the proposed models and algorithms on a real Tencent Weibo dataset containing 2.33 M users, 51 M links and 6 k different topics.

The rest of this paper is organized as follows. Section II provides the preliminary background and highlights the related work. In section III, we formulate the problem statement, whereas section IV contains our proposed solution framework and the maven model. Section V presents the algorithms for maximizing the information flow in the mavens’ social graph. Evaluation results based on real and synthetic data sets are shown in section VI. Section VII concludes our paper.

II. RELATED WORK

To illustrate our approach, we will highlight the topics that are thoroughly discussed, which include influence maximization and mavens in the marketing field.

A. Influence Maximization Problem

Influence maximization has been extensively studied in the literature. In 2003, Kemp et al. [23] formulated influence maximization as an optimization problem. They proved the NP-hardness and sub-modularity of influence maximization under the two models presented in their work. They used a greedy algorithm as an approximate solution to this problem. The greedy approximation had already been proved which was (1-1/e) approximation on monotonic and sub-modular functions [22]. Chen [11] proposed efficient algorithms and heuristics for the influence maximization problem. Moreover, a topic influence model was studied in [2], and [7] to add a topic-modeling perspective of social influence by introducing topic-aware propagation models and identifying influencers in topic-specific networks. In [2], the authors introduced a new influence propagation model instead of considering user-to-user influence. In [2], the model considers users’ authoritativeness and interest in a specific topic and how an item is relevant to a specific topic. Then [7] proposed a general search framework for finding topic-specific key influencers with various models. In contrast, we show in this paper that seed selection in social networks depends more on personal characteristics than on the item’s topic or category, especially in viral marketing applications. On the other hand, [8] highlighted some key people in social networks and attempted to define their personal characteristics through their activities. However, the definitions in this work are not extendable to microblogs like Twitter and Facebook. It is worth mentioning that none of these works developed an information cascade model under marketing characteristics identifications.

B. Mavens in Business Field

The first illustration of the market maver concept was in 1987 by Feick and Price [13]. The term mavens basically refers to a group of consumers that enjoy generalized marketplace information and takes a strong interest in broadcasting this to others. Thus, mavens act as a critical means of spreading product information, often with greater credibility than many
traditional marketing communications sources. In marketing, the existence of the market maven has prompted extensive research for both physical channels (i.e., real-world) [1], [13] and the modern web-based channels [4],[5]. As a part of our research, we focus on identifying mavens in a given social network based on their activities. To accomplish that, it is worth summarizing the most significant characteristics of mavens in social networks. [9] summarized the personality characteristics of mavens and clearly emphasized the difference between mavens and other special users like opinion leaders and early adopters. Briefly, mavens tend to have general and multiple interests which typically “contrast with opinion leaders and early adopters, who are more knowledgeable and want to share information on specific ranges of products within a product category or specific market environment characteristics”. Moreover, [4] illustrated the main personal characteristics of mavens such as high communication accompanied with media consumption about multiple products compared to others, a positive attitude toward advertising and an awareness about consumption about multiple products compared to others, a product category or specific market environment characteristics”. Moreover, [4] displayed and tested the main characteristics that are observable in the virtual world and social networks. These characteristics include taking pleasure in sharing information to reinforce the maven’s image in their community, spanning multiple product categories making them a useful target for companies, and a high degree of individualism. To summarize, the influence maximization problem is an optimization problem that aims to identify seed nodes that can successfully generate the maximum cascade, and the current research lacks utilizing marketing identifications to enhance the selection of seeds.

III. PROBLEM DEFINITION

In a social network, we are given a social graph in the form of $G = (V,E)$, where the nodes $V$ are users. A directed edge $(u, v) \in E$ between users $u$ and $v$ represents a social relation initiated by a user $u$ toward a user $v$. We will refer to neighbor nodes that are connected to a node $u$ by $N(u)$. In addition, we have the users’ action log which contains every action performed by every user of the system. The users’ action log consists of tuples in the format $(u, a_u, t_u)$, which indicates that a user $u$ performed an action $a_u$ at time $t_u$. Based on this action log, we assume that the set of nodes $V$ of the social graph $G$ is extracted from the first column of the action log. On the other hand, let $A$ denote the universe of actions where each action is represented as a vector containing its name and some additional features for actions. For instance, some actions contain tag people, pages, or adding locations. Each action has the format $a_i = [a_{i, property(1)}, \ldots, a_{i, property(n)}]$. We let $A_u$ denotes all of the actions performed by a user $u$. In the following, we introduce some definitions in order to capture the main characteristics of mavens:

**Definition 1: (User’s individualism)** Let $A_u^*$ refer to actions that have been performed by a user $u$ before all his connected users in the social network. Thus, $A_u^*$ can be written as: $A_u^* = \{a_i | \forall v \in N(u), if \exists (u, a_i, t_u), (v, a_i, t_v) \in \text{ActionLog}, t_u < t_v\}$. In addition, let $A_u^-$ represent actions that are performed by only a user $u$ among its neighbors $N(u)$. $A_u^-$ will be defined as follows: $A_u^- = \{a_i | \forall v \in N(u), (u, a_i, t_u) \in \text{ActionLog}, (v, a_i, t_v) \notin \text{ActionLog}\}$. The user individualism $\text{ind}_u$ can be measured as follows:

$$\text{ind}_u = \frac{|A_u^* \cup A_u^-|}{|A_u|}$$  \hspace{1cm} (1)

**Definition 2: (User’s sharing desire)** Let $\hat{A}_{u,v}$ denote the actions of a user $u$, where this action contains a property to include another user $v$, $\hat{A}_{u,v}$ is expressed as follows: $\hat{A}_{u,v} = \{a_i | \exists v \in N(u), (u, a_i, t_u) \in \text{ActionLog} : a_i = [a_{i, property(1)}, \ldots, a_{i, property(n)} = v]\}$. The users sharing desire $\text{shr}_u$ can be measured as follows:

$$\text{shr}_u = \frac{\sum_{v \in N(u)} |\hat{A}_{u,v}|}{|A_u|}$$  \hspace{1cm} (2)

**Definition 3: (User’s multiplicity of spanning )** Let $\text{ Topic}_u$ represent all topic categories in the given action log, and lets $\text{ Topic}_u$ denotes the topics categories that have been included in $A_u$. Thus, to measure a user’s spanning of multiple product classes can be defined as follows:

$$\text{mul}_u = \frac{|\text{ Topic}_u|}{|\text{ Topic}|}$$  \hspace{1cm} (3)

After defining the mavens main characteristics in social networks, we create for each user $u$ a vector of data $[\text{ind}_u, \text{shr}_u, \text{mul}_u]$ which represents a feature vector to summarize the maveness confidence level for a user $u$. This leads to redefining the social graph based on the maveness confidence level.

**Definition 4: (Mavens graph )** For a given Social Graph $G = (V,E)$, and corresponding action log $(u, a_u, t_u)$, we create a weighted mavens graph $G = (V,E,M,D)$, where $M : [\text{ind}_u, \text{shr}_u, \text{mul}_u] \rightarrow R$ is a function that calculates the maveness confidence level for each user, and $D : E \rightarrow R$ is a diffusion frequency function that represents the frequency of contacts from user $u$ to a user $v$ as follows:

$$D_{(u,v)} = \frac{|\hat{A}_{u,v}|}{\Delta}$$  \hspace{1cm} (4)

where $\Delta$ is a specific period of time and $|\hat{A}_{u,v}|$ is the number of actions performed by a user $u$ toward a user $v$ in the determined period $\Delta$.

The maven graph consists of users with their conference level of being mavens, and edges represent the frequency of diffusion or the strength of the social tie that connect them in the direction of the edge. In this graph, we want to find the maximum multi-commodity flow in the network where the capacity of each outgoing edge does not exceed the maveness confidence level of the user. In mavens graph, the nodes that have a high maveness confidence level will be considered as sources, and each of them will be paired with all remaining nodes which are considered sinks. The problem we tackle is to find the maximum flows that satisfy the node conservation constraints so that the sum of flows on
any edge does not exceed the capacity of the edge. To achieve this, we first need to learn a maveness confidence function $M : [\text{ind}_{u}, \text{shr}_{u}, \text{mul}_{u}] \rightarrow R$ for each node in the graph. After that, we will select source nodes by only considering nodes with a high maveness level, i.e., $Mav = v : m_{v} \geq \Omega$, where $\Omega$ is the maveness confidence threshold value which will be chosen later, and the remaining nodes are sinks. Therefore, we will derive the flow function in the social network through mavens. In particular, we want to derive a flow function $f(v, u)$ that cannot exceed the capacity or the maveness confidence of the node that sends the message $f(v, u) \leq m_{v}$. Ultimately, we want to maximize the multi-commodity flow from mavens to the whole network such that the sum of the flows of all commodities is maximized. Formally our problem is defined as follows:

$$
\max \sum_{v \in MAV} f(v, u)
\text{s.t.} \sum_{v \in MAV} f(v, u) \leq m_{v}, f(v, u) \geq 0.
$$

where $f(v, u)$ is the flow function from a maven node $v$ and $m_{v}$ is the maveness confidence level of a node $v$.

IV. SOLUTION

For our solution’s framework, in order to derive the flow function through mavens, we adopt an information cascades model. In particular, we used Bayes’ rule to develop a model of decision-making under uncertainty. In this model, we calculate the probability of any particular user $u$ of adopting or being influenced by a certain message or a product by using the defined characteristics to reason about decision-making. For instance, if a user $u$ is active and has a maveness confidence level $m_{u}$, there is a social connection with an inactive user $v$ who has a maveness confidence level $m_{v}$, and the diffusion frequency from $v$ to $u$ is $d_{v,u}$, the probability of $v$ being influenced is

$$
Pr(v_{\text{active}}|u_{\text{active}}) = \frac{Pr(v_{\text{active}})Pr(u_{\text{active}}|v_{\text{active}})}{Pr(u_{\text{active}})} = m_{v}m_{u}d_{v,u}
$$

To generalize this model, if a user $v$ has multiple neighbors where some of them are active and the rest are inactive, we will denote their states as $U = \{U_{\text{active}}, U_{\text{inactive}}\}$ to set multiple neighbors $U$ spreading information to a user $v$ are independent of each other. Thus, the joint probability $Pr(v_{\text{active}}|U_{\text{active}})$ will be calculated as follows:

$$
Pr(v_{\text{active}}|U_{\text{active}}) = \sum_{i=1}^{[U]} Pr(v_{\text{active}}|U, U_{\text{active}}).Pr(U|U_{\text{active}})
$$

Subsequently, we have to ensure that the proposed model satisfies the occurrence of a fully revealing informational cascade with any given priors [20]. Thus, we introduce the following theorem.

**Theorem 1:** Information diffusion through mavens has a fully revealing informational cascade with probability 1 for all prior diffusion frequency $d$ if the number of activated users $U$ before user $v$ overcomes the number of rejections by two or more.

**Proof:**
The proof starts by introducing $Pr(v_{\text{active}})$ to represent the probability of a user $v$ positively responding or being active for some message $s$ and the states are either active or inactive. Based on the crowd phenomenon [10], we will get

$$
Pr(v_{\text{inactive}}) \leq Pr(v_{\text{active}})
$$

using Lemma 2 in [20] and inequality 8, we will define $\varepsilon'$ and $\varepsilon''$ as follows:

$$
m_{v} - \varepsilon' \leq Pr(v_{\text{inactive}}) \leq Pr(v_{\text{active}}) \leq m_{v} + \varepsilon''
$$

With this assumption, either $\varepsilon' = 0$ or $\varepsilon'' = 0$, we will rewrite this inequality considering the prior probability of mavens and the diffusion frequency as follows:

$$
\frac{\sum_{u \in U} d_{u,v}(1-m_{u})}{\sum_{u \in U} d_{u,v}} \leq m_{v} + \varepsilon''
$$

To satisfy this inequality, we should find the solution of the following system:

$$
d_{u,v} \geq 0 \forall (u, v) \in E
$$

$$
\sum_{u \in U} d_{u,v}(1-m_{u})(-m_{v} + \varepsilon') \geq 0
$$

$$
\sum_{u \in U} d_{u,v}m_{u}(-m_{v} - \varepsilon'') \leq 0
$$

By applying Frkas lemma [12] to solve these linear inequalities, we will use $\kappa_{u}$ for all $u \in U$ as a dual variable for $d_{u,v}$ to obtain the following equations:

$$
\kappa_{u} + (1-m_{u})(-m_{v} + \varepsilon')\kappa = d_{u,v}m_{u}(-m_{v} - \varepsilon'') \forall (u, v) \in E
$$

Since $\frac{m_{u}(-m_{v} + \varepsilon')}{(1-m_{u})(-m_{v} - \varepsilon'')} = \frac{1}{\kappa_{u}}$ is increasing in $m_{u}$, there is a $\kappa$ that satisfies the system, where

$$
\frac{Pr(v_{\text{active}})(Pr(v_{\text{active}})-m_{v}-\varepsilon')}{Pr(v_{\text{active}})(Pr(v_{\text{active}})+m_{v}+\varepsilon'')} \leq \frac{Pr(v_{\text{inactive}})(Pr(v_{\text{inactive}})-m_{v}-\varepsilon')}{Pr(v_{\text{inactive}})(Pr(v_{\text{inactive}})+m_{v}+\varepsilon'')}
$$

which satisfies the necessary and sufficient condition for a fully revealing informational cascade based on theorem 2 in [20].

Based on the above model, we can define the flow function from a node $u$ to a neighbor node $v$ as follows:

$$
f(u,v) = \frac{m_{u}m_{v}d_{u,v}}{d_{u,v}}
$$

We use this function to maximize the multi-commodity flow in the mavens graph, where source nodes are the nodes with high maveness confidence levels. In fact, we will treat the problem as packing s-t paths so that the constraints imposed by the maveness confidence level and diffusion frequency are not violated. To solve this problem, we follow the GargKonemann approach [19] by associating a length value with each edge. At any step $i$, we select a unit flow along with the shortest s-t path. Then we update the distance of every edge on this path
by $1 + e$ for a fixed $e$. By applying this, we guarantee that we always choose the shortest s-t path to route flow along. Thereby, the flow is balanced on all edges in the graph.

This model leads to the necessity to calculate the maveness confidence for each user based on their personal characteristics.

A. Learning Personalized Mavens Characteristics

As a result of the users’ characteristics measurements, we can create for each user $u$ a vector of data $[\text{ind}_u, \text{shr}_u, \text{div}_u]$ which represents a feature vector to summarize the measurements of the main characteristics of mavens. However, the proper weights for these parameters to classify mavens are unknown. Because both feature vectors, in addition to class labels can be used to estimate the model that describes the classes (and a totally arbitrary model is difficult to handle), some assumptions have to be made about the structure of the estimating model. Accordingly, the classification of unknown samples is based on estimated class representations in a feature space. The suggested classifier model is the Gaussian mixture model (GMM) to represent the feature space because the feature has an approximately normal shape in density distribution, as shown in Fig. 1, which makes it suitable for a class model in the feature space [21]. Basically, a Gaussian mixture model is a weighted sum of given feature components that have Gaussian densities, as given by the equation

$$E(x|\theta) = w.g(x|\mu, \sigma)$$

where $x$ represents the number of the measured features in the feature vector, while $w$ represents the mixture target weights, and $g(x|\mu, \sigma)$ are Gaussian densities of the components where vector $\mu$ is the mean value and $\sigma$ is the covariance matrix. Each component density can be obtained by the following equation,

$$g(x|\mu, \sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu) \right\}$$

considering $\theta$ includes the values of $w, \mu$ ,and $\sigma$.

It should be taken into the account that the only constraint on the sum of the weights for features vector is satisfying $\sum w = 1$.

In details, our target is to calculate the parameters that give us a classification of users or to select an initial model that fits the observed data and ensure that the data likelihood has such a goodness value.

The process of estimating the maximum likelihood of the weights needs to be determined using an iterative method such as the expectation maximization (EM) algorithm [21] to calculate the likelihood function:

$$\ell(X|\theta) = \prod_{n=1}^{N} E(x_n, \theta)$$

where $X$ is a set of independent samples $X = \{x_1,...,x_N\}$ used by a probability density function and the objective is to find $\theta$ that maximizes the likelihood:

$$\theta_{opt} = \max \ell(X, \theta)$$

The purpose of using the expectation maximization (EM) algorithm is that the above function is a non-linear function of the parameters $\theta$. Thus, it is not possible to obtain a direct maximization. In the EM algorithm we will initially begin with a random guess for $\theta$, which leads to get a new value $\hat{\theta}$, that satisfies $\ell(X|\hat{\theta}) > \ell(X|\theta)$ The iteration will be repeated until an acceptable threshold value of convergence is reached. Eventually, the mixture weights will be:

$$\overline{w}_i = \frac{1}{N} \sum_{n=1}^{N} p(i|x_n, \theta)$$

V. Algorithms

This section illustrates the algorithms we developed in our solution. We start by learning the maveness confidence level in algorithm 1 using(EM). The expectation step (E-step) (lines 1-6) implements equation 14, where $\theta_i$ is the previous estimate of the distribution parameters, and $\theta$ is a variable for a new estimate that describes the (full) distribution. Precisely, $L$ in equation 14 calculates the likelihood of the data, considering the unknown class with respect to the current estimate of the distribution described by $\theta_i$. The M-step (lines 8-17) is used to maximize $Q(\theta; \theta_i)$ with respect to equation 17. The steps are repeated until a convergence criterion is met (line 15). Finally, the maveness confidence level for each user is calculated based on the selected $\theta$ (line 18). The time complexity of algorithm 1 is $O(3|V| + 2|V|^2)$.

Algorithm 2 creates the maven graph by eliminating the nodes that are not mavens and do not have a path to connect them to any maven. Initially, the algorithm assigns empty set for all $V'$, $E'$, and $P$. Line 3 illustrates that the only nodes will be added to the mavens graph are maven nodes and their neighbors. We calculate the length function $l(u, v)$ for the edge between node $u$ and $v$ in line 6. Line 7 shows that the only edges that will be added to $E'$ and $P$ are edges with a positive
Initially, we assign the flow in the network to zero, and the initial value of α is calculated using the following function:

\[ Q(\Theta, \Theta_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2}(x-\mu)^2 \right\} \]

for any \( i \). The output of Algorithm 2 is the maveness confidence level of each user, which is then used to calculate the maveness confidence level of the node set selected in Algorithm 3. The time complexity of Algorithm 2 is \( O(|V|^2) \).

Turning to Algorithm 3, which aims to maximize the multi-commodity flow in the network using the following procedure. Initially, we assign the flow in the network to zero, and the length of the shortest path between nodes \( u \) and \( v \) as the distance between them (line 2). Line 2 also clarifies that in each iteration, we will select the path that has the shortest distance value. The flow in this path will be determined by the smallest maveness level of the nodes that create the selected path (line 4). The network flow will be updated in line 5 and the distance value for each edge in the selected path will be updated in line 6.

Algorithm 3: Maximum Multi-commodity Flow Finding

```
Input: \( G=(V',E') \)
Output: max flow \( f \), min length \( l \)
initialization: dist(\( u,v \))=min \( l(\( u,v \)) \) ; \( f=0 \)
for \( p(\( u,v \)) \in P \) do
    \( p_{\text{selected}} \leftarrow \min \) dist(\( u,v \))
    while dist(\( u,v \)) < 1 do
        \( m \leftarrow m_i \) in \( p_{\text{selected}} \)
        \( f = f + m \)
        for \( (u,v) \) in \( p_{\text{selected}} \) do
            dist(\( u,v \)) = dist(\( u,v \)) + \( 1 + \frac{m}{m_{\text{max}}(u,v)} \)
        end
        \( p_{\text{selected}} \leftarrow \min \) dist(\( u,v \))
    end
end
Return f, dist ;
```

Ultimately, algorithm 4 identifies the k-node set based on the maven graph. The k-node set is selected as follows. First, we mark \( V_{\text{temp}} \) as a copy of \( V' \), and then choose a node \( u \) from \( V' \), that can maximize the flow among the remaining nodes (line 2). After that, we omit the node \( u \) and all its successors from the graph (line 4). This process is executed iteratively until either the number of k seeds is obtained or all network nodes are activated(line 1). The time complexity of algorithm 4 is \( O(k|V'|^2) \).

Algorithm 4: Identifying the k-node Set

```
Input: \( V',k \)
Output: S
initialization: \( V_{\text{temp}} = V' \)
while \( |S| < k \) and \( V_{\text{temp}} \neq \emptyset \) do
    \( u = \max \{ f(\( u \in V' \cap V_{\text{temp}} \) \}
    \( V_{\text{temp}} = V_{\text{temp}} \setminus \{ v | v \in p(u,x) \} \) \)
end
Return S ;
```
VI. EXPERIMENT

Our experiments have two goals. In the first one, we want to detect mavens in a social network and evaluate their effect in reshaping the social network graph. The other objective, is examining whether mavens can play a better role than influencers in a social network to maximize the spread of word of mouth. The expected result of using the mavens model is enhancing the social graph in terms of reducing the nodes based on a deeper analysis of users’ characteristics rather than dealing with all of the users and treating them based only on overall actions which will significantly improve the efficiency and speed of the resulting social graph model. We also aim to consider whether the effect of mavens can be considered to be an alternative to influencers, bearing in mind that the mavens’ role is more accurate and precise when measured in a social network. Therefore, we compare the maximization flow in a mavens graph against the topic aware model AIR in the influence maximization problem.

Datasets and experiment setup: In our experiment, we used a dataset from a famous Chinese microblog site ?Tencent Weibo (t.qq.com). This dataset is released by KDD Cup 2012. This dataset includes 2.33 M users and 51 M links. The total amount of words used is 492 M distributed among 6 k topic categories. Generally, the dataset represents users actions like the recommendation of items, along with profiles of users’ “follow” histories. Each user in the dataset is associated with rich information, i.e., follow history, profile keywords, and items recommendations with their time stamps. We conducted the experiments on an Intel(R)Core(TM)i7-4510U 2.6 GHz CPU machine with 8 GB RAM.

![Fig. 2. ROC analysis](image)

**Fig. 2. ROC analysis**

Accuracy of Learning: We started by learning the users’ characteristics. In fact, the used dataset provides the action log in several distributed tables. Therefore it was necessary to derive the needed parameters from the given tables. For instance, to derive the individualism of a user \( u \), we had to refine the actions where user \( u \) was either the only one or the early one. To obtain that, we relied on a user-keywords table to refine the words that have been used by a user \( u \) and have not been used by any friend to calculate the unique actions. For the early actions, we used the recommendation log table to count all the attempts of a user \( u \) to activate any neighbor even unsuccessful attempts. In order to learn the confidence level of maveness, we split the dataset based only on the recommendation actions table, such that a user action can appear either in the training or test dataset. In order to evaluate our predictive accuracy, we compared our mavens model with the maven model in [8] by means of ROC curves. Each point in the ROC curve corresponds to \( \Omega = .8 \), which is the same for all users. The purpose of this test is to measure whether our maveness confidence level can predict the amount of overall user recommendation actions. This is basically a binary prediction task: for a given maveness confidence level in the training set, we try to predict the amount of actions in the testing data set without considering the timestamp in this test. On the other hand, we used the maven’s definition in [8] to test the same amount of actions. Figure 2 illustrates that our definition for a maven performed better in estimating users’ action behavior than the definition in [8]. Hence, the definition in [8] only highlights the top influencer users, and the test allows us to evaluate the contribution of maven characteristics modeling to the prediction of users’ actions. The impact of the mavens in the graph: In this experiment, we evaluate the maveness concept in reducing the nodes number in the mavens graph using the proposed algorithm 2 based on different \( \Omega \) threshold values. As shown in Figure 2, the number of nodes in the resulting graph is dramatically decreased compared to the number of nodes in the original social graph. In particular, we tested several threshold values to evaluate the impact of the size on the resulted social graph. For instance, a 0.65 maveness confidence level will lead to a sharp increase in the node size of the mavens graph. In contrast, assigning 0.9 as the maveness confidence level could be deceptive and could lead to excluding a set of nodes that might be valuable. Specifically, we found that \( \Omega = 0.8 \) is the optimal threshold value because the size of the resulting nodes uniformly increases counter to the original social graph node size. Moreover, applying this method will also control the size of the mavens graph to be investigated. It is worth mentioning that we did not find any relation that determines the optimal maven graph size based on the seed size. Thus, this point will be considered in future work. Influence Maximization vs. Mavens Flow Maximization: In our experiment, we perform a comparison between AIR social influence propagation and mavens diffusion. We specifically selected the AIR model because it was the reference that...

1 www.kddcup2012.org/c/kddcup2012-track1/data
highlighted the importance of introducing a new model based on some users characteristics instead of considering the user-to-user influence. The experiment was implemented for AIR by selecting 50 different random items with their associated categories and calculating the expected spread using 1000 Monte Carlo simulations. Alternatively, we ran Algorithm 4 on the resulted mavens graph. In Fig 3, we summarize the expected diffusion achieved by k-seed influencers on the AIR propagation model in front of the expected cascade that resulted from k-seed mavens on the mavens graph. Indeed, the mavens greatly exceed influencers in spreading the words in the social network. In the used dataset, the top maven node successfully recorded a spread among 88 nodes against only 27 nodes affected by one influencer. In addition, the mavens flow reached about 800 nodes using only 25 mavens. In the opposite direction, just 130 nodes were influenced by 25 nodes in the AIR model. In summary, mavens achieved the best performance in effectively reducing the node size in the social graph and the maximum information spread in Tencent Weibo.

VII. CONCLUSION

This paper introduced mavens in social networks. We presented a model to detect the main characteristics of mavens. We applied our model in Tencent Weibo and verified that our methods detect mavens and provide a good estimation of their behavior in the network. We also studied a way to maximize information flow through mavens in a social network. Our experiments rivaled that distinguishing the nodes based on their maveness confidence will improve the efficiency of the resulting social graph by rapidly reducing the nodes size. We also emphasized that mavens widely maximize the information flow in a social network compared to the limited effect of influencers in the influence maximization problem. Therefore, we are looking to extend our work in this paper. First, we would like to investigate the relation between the desired seed size and the node size in the resulting mavens graph. In addition, we would like to explore how to combine influence maximization with mavens modeling to increase the robustness of social graph modeling.

REFERENCES