12-14-2016

Identifying Mavens in Social Networks

Hussah Albinali

IDENTIFYING MAVENS IN SOCIAL NETWORKS

by

HUSSAH ALBINALI

Under the Direction of Dr. Yingshu Li

ABSTRACT

This thesis 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 develop some methods to learn their characteristics based on a real dataset. We also illustrate the ways to maximize information flow through mavens in social networks. Our experiments show that our model can successfully detect mavens as well as fulfill significant roles in maximizing the information flow in a social network where mavens considerably outperform general influencial users for influence maximization. The results verify the compatibility of our model with real marketing applications.

INDEX WORDS: Social Network, Influence Maximization, Algorithm.

IDENTIFYING MAVENS IN SOCIAL NETWORKS

by

HUSSAH ALBINALI

A Thesis Submitted in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Computer Science

in the College of Arts and Sciences

Georgia State University

2016

IDENTIFYING MAVENS IN SOCIAL NETWORKS

by

HUSSAH ALBINALI

COMMITTEE CHAIR: Dr. Yingshu Li

COMMITTEE : Dr. Wei Li

Dr. Zhipeng Cai

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

December 2016

ACKNOWLEDGMENT

I am writing this note of thanks in the finishing touch on my master thesis. It was an intense learning period for me, not only in the research area but also on writing and communication level. Writing my thesis has had a big impact on my knowledge and skills. I would like to thank the people who supported me throughout this experience. First of all, I would like to thank my thesis advisor Dr. Yingshu Li of Computer Science Department at Georgia State University. She always enthuses whenever I ran into a trouble spot or had inquiries about my research. I would also like to thank Meng Han who is a Ph.D. student in Computer Science Department at Georgia State University. I am gratefully appreciative to his valuable feedback on this thesis. Finally, I must express my thanks to my family for the support that they provided through my study journey. Without their love, encouragement and believe, I would not have finished this thesis.

Hussah Albinali

Table of Contents

List of Figures

List of Tables

1. 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 [21],[9]. A viral marketing technique is essentially based on a network model where users exchange their knowledge and experiences throughout preexisting social relationships. Kempe Goyal et al. [9] 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 Linear Threshold Model or the Independent Cascade model is NP-hard. 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 [22] 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 [5]. In fact, those classes are very close to high potential consumers in viral marketing from the marketing perspective which involves market mavens, social hubs, and salespeople [19]. 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 influence maximization problem. As mentioned above, in [22] 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. 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" [16]. This group was first introduced by Feick and Price [16]. The existence of market mavens

later received widespread attention in physical as well as web-based channels [29]. 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 trying 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 avoid the most serious pitfalls of failing to ignite the influence seed. This research 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 thesis is organized as follows. Section 2 provides the preliminary background and highlights the related work. In section 3, we formulate the problem statement, whereas section 4 contains our proposed solution framework and the maven model. Section 5 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 6. Section 7 concludes the thesis.

2. RELATED WORKS

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

2.1. Social Networks

Social network sites (SNSs) like Facebook, Weibo, and Cyworld have attract-ed lots of users in a way that these sites become an essential part of their daily routines. As these social network sites support a various aspect of in-terests and practices, however, most of the features in these sites are fairly consistent.
Social network sites also vary in utilizing and incorporating new communication means, such as mobile connectivity, locations detection, and photo/video-sharing. This vivid development in mobile social networks leads scholars to study the implications of these sites, as well as analyze users engagement with them. First, we would like to introduce a definition for a social network. Following this, we contextualize major users' roles on SNSs and attempt to highlight key works.
We define social network sites as a web mean that allows users to create an optional public profile within a limited system, in addition to articulate a list of other users where they share connections. Social network sites also enable

users to view and explore their list of connections and possibly other users within the system. The nature of these connections may differ from site to site. It is worthwhile to mention a distinguishing feature of the social network which is the possibility of individuals to meet strangers by allowing users to spotlight and make visible their social networks. This feature leads to the existence of connections between individuals that would not then be made. Nonetheless, that is mainly not the goal, and these relations are frequently between"latent ties" who share some real connection. On some large SNSs, users are not essentially "socializing" or looking to know new friends; instead, they are mainly communicating with their original mates of their extended social network.

While SNSs reign a wide variation of technical features, their pillar consists of users' profiles that display a list of friends who are obviously other users of the system. In fact, Profiles are unique pages where users can express themselves. After joining an SNS, a user is asked to fill out some forms that contain a series of questions. The profile is created using the answers about me questions, which typically include descriptors such as age, location, and interests. Most sites also offer a profile photo feature. After joining a specific social network, users are trying to recognize others in the system with whom they have a common interest or a real connection. Even though the label of relationships may differ on the sites including friends, contacts, and fans. De-

spite the fact that these labels can be confusing, these labels basically refer to the existence of a connection between the relationship parties and does not necessarily mean friendship in the regular everyday sense. Generally, SNSs require bi-directional approval for friendship. However, this is not always the case. In fact, one-direction of relationship also exists and this kind of ties is sometimes called followers, or fans. The availability to explore public connections is a vital element of SNSs. Each friend has a list contains links to each Friends profile, which enables viewers to traverse the network users by clicking through the Friends lists, where this list is visible to any user has a permission to view the profile. SNSs largely vary in the available features and user base. Additional features in most SNSs include comments and private messaging. Some have photo-sharing or video-sharing capabilities; others have mobile-specific SNSs as well as supporting mobile interactions for web-based SNSs like Facebook, MySpace, and Cyworld.

Turning to SNSs roles to bridge online and offline social networks, the research suggests that most SNSs primarily advocate current social relations. For instance, [23] found that Facebook is used to preserve existing offline relationships or solidify real connections, as contrasting to creating new relations where these relationships suffer from weakness. [23] also investigated the interactions between online versus offline ones. This ability of SNSs in creating a social environment make them deeply embedded in users lives. For instance,

Cyworld has become a vital part of everyday life in Korea. [12] found that %85 of that study respondents "listed the maintenance and reinforcement of pre-existing social networks as their main motive for Cyworld use". Similarly, [23] confirmed that and Facebook allows U.S. youth to meet their friends even when they are unable to socialize frequently. She also states that SNSs are "networked publics that support sociability, just as unmediated public spaces do".

Emphasizing on network structure, social network sites sturdily provide a rich source of behavioral data. Profile and connections through SNSs datasets allow researchers to explore patterns of relationships, behavior, and even further analyses for trends indicators. This investigation started with examinations of blogs and other forum websites. To illustrate that, [23] examined a massive dataset consisting of 4 million users exchanged 362 million messages on Facebook seeking to analyze friending and messaging activities. SNSs researchers have also studied the network structure of friendship analyzing the roles that people played in the growth of Social Networks. For instance, scholarship concerning LiveJournals network has included a friendship classification scheme [15], with analysis of the role of language in the topology of Friendship [10]. Turning to the inquiry about hot to detect patterns and identify existing communities in social networks or in other words the projections of the network

on a particular or a specific actor who have all the neighbors at some distances which are defined a priori or influencer. To clarify this point, it is important to differentiate between the notions of "role" and "position", in social networks. The roles can be defined as specific behaviors and interactions which can be observed (e.g., giving orders, sending emails, etc.). While a position means a well-defined place in a social structure like for instance, 'parent' and 'child' are both eligible positions in the network. According to these definitions, positions and roles together can identify social relations. In [20] found that actors with similar roles will share common features and common patterns of relations, even if they do not share any direct relationship. In order to detect non-explicit roles in social networks, we need some machine learning methods to the perspective of dealing with data that are relational and they violate exchangeability assumptions and the regular independence. Usually to achieve this many assumptions will be made based on some dependent observations which are consequences of enormous connected data. These methods include clustering algorithms using the graph structure. The content of any exchanged documents, like emails or posts, is also used. To that end, the available background including both the exchanged text along with social relations are used in the definition of the roles based on the assumption that the roles will result from the regularities that are found in the structure of the social relations and the features describing users. Besides role detection, identifying positions is

achieved by using block models, which is an algebraic framework that deals with various issues of social networks such as the identification of communities and their in-between relations, should handling multiple relations, in this case, involves dealing with multigraphs. Another aspect to consider in identifying social relationships is estimating roles using probabilistic models on textual content. While in [20] the authors argue that the relational structure is not enough when analyzing textual datasets, such as emails, blogs, scientific papers. The use the textual content should be associated with relational structures i.e., the social graph should be predefined, and any document like email, post, or message is only supporting to identify the roles.

In addition, [23] claimed that friend connections are not the only network structure worth investigating. They studied the ways in which the performance of media i.e. favorite music, books, or film will constitute an alternate network structure, or as they call it a taste fabric. This property particularly attracts many marketing companies and organizations to promote their products. Thus leads to define a major problem that is to be considered in this domain is the influence maximization problem.

To represent the context of personal relations between internet users, the social graph term was introduced in 2007 at the Facebook F8 conference [7]. The social graph term was used to explain the way that the newly introduced

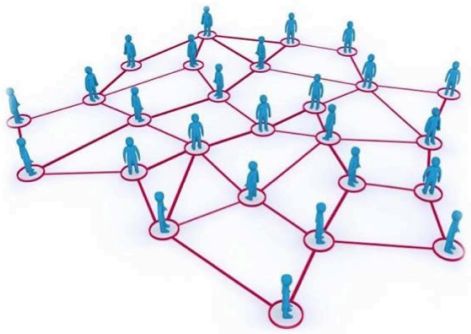

Fig. 2.1: Social Networks Graph

by Facebook Platform would benefits of the relationships between individuals to offer a richer online experience. Then the definition is expanded to refer to a social graph of all social networks sites users. Briefly, this model reshapes a social network to support a mathematical analysis for the relational representation in a social network. After explaining the concept of the social graph, Facebook has often touted by offering the social graph platform pen Graph API to other websites so that users' relationships can be collected to use on websites outside of Facebook's control.[7]

2.2. Influence Maximization Problem

The social network is the largest communication medium that connects millions of people to share their ideas and other information. Social networks act as a channel for information distribution. Viral marketing application and influence maximization problem in social networks are tightly related to each

other. This section will highlight remarkable works in the influential maximization in a large scale social network. It will also explain many effective algorithms; also it will point out some significant problems such as scalability, accuracy, and growth of the network. Influence maximization has been extensively studied in the literature. As previously mentioned, the goal of influence maximization is to identify a small subset of individuals in a social network that they early adopt a new product or opinion and generate a large cascade of individuals for the adoption process in the network. In practical, influence maximization problem has many applications including a viral marketing approach or word-of-mouth marketing where users grab more customers by only sharing their experiences and disseminates information in a network. Many organizations, in fact, make use of this strategy to promote their products, and innovations through a social network. Thus, influence maximization problem tackles selecting some nodes as a starting step to spread information to a huge amount of nodes in a network.

The problem of influence maximization was firstly introduced by Domingos and Richardson in [25] as it introduced a descriptive model and used a joint distribution in nodes' behavior. However, Kempe et al. [9] has proved that the influence maximization problem is NP-hard. They also were the first to formulate influence maximization as an optimization problem. Their work will be illustrated as follows. The social network is represented by a directed graph

of relationship where the inputs of this problem are the data on a social network, estimation for the extent to which individuals influence one another, and a new product that will be adopted by a large fraction of the network. The expected outputs are the top key influential people as well as the total interested people. The problem mainly is how to choose the best key people to get the maximum interested people. Giving more detail about the input network data are users which represent the vertices of the graph. Each user has a state either to be active which means an adopter of the innovation or inactive. It is worthwhile to mention that the activities in the users states increases monotonically by users' neighbors. They proved the NP-hardness and sub-modularity of influence maximization under the two models presented in their work. The two basic diffusion models, namely Linear Threshold (LT) and Independent Cascade (IC) provide a platform to achieve the influential maximization. In Linear Threshold Model a node v is influenced by a neighbor w with a weight $b_{v,w}$ such $\sum_{neighborsofw} b_{v,w} \leq 1$. In addition, there is θ_v random value from the interval [0,1] considering the existence of A_0 an initial set of active nodes. The diffusion process in this model works as following, in step t, all nodes that were active in step $t-1$ remain active. Furthermore, any node v for which the total weight of its active neighbors is at least θ_v will be also activated. In another word, $\sum_{neighborsofw} b_{v,w} \geq \theta_v$. On the other hand, the process in Independent Cascade Model of unfolds runs in discrete steps. First,

when a node v becomes active in step t, it has only a single chance to activate each currently inactive neighbor w. The probability of success activation $p_{v,w}$ is a predefined parameter. If w has several newly activated neighbors, their activation attempts are ordered sequentially. If v succeeds, then w will become active in step $t+1$. Certainly, the process runs until no more activations are possible. 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 [11].

For influence maximization problem, proofing the submodularity in the influence function for independent cascade model as well as the linear threshold model was a base to many other contributions. It is worthy to highlight the fundamental of these proofs. The submodularity condition states if S and T are two sets of nodes where $S \subseteq T$, $\sigma_x(S \bigcup \{v\}) - \sigma_x(S) > \sigma_x(T \bigcup \{v\}) - \sigma_x(T)$. For independent cascade model, the influence function $\sigma(.)$ is based on the claim that any node x ends up active if and only if there is a path from any active node in the active node set A to x consisting entirely of live edges. Following the probability space rules where each sample point specifies one possible set of outcomes for all the coin flips on the edges. By considering X one sample point of the outcomes $\sigma X(A)$, the expected number of active nodes by the activation process when A is the initially targeted set. Because we have determined a choice for X, $\sigma X(A)$ is a deterministic quantity. By letting $R(v, X)$

to represent the set nodes that can construct a path from v consisting the entire live edges. $\sigma X(A)$ is the number of nodes that are included in live-edge paths from any node in A, and so it is equal to the cardinality of the union $U_{v \in A} R(v, X)$. The number of elements in $R(v, X)$ which are not already in the union $U_{u \in SR} R(u, X)$ can be considered as $\sigma_x(S \bigcup \{v\}) - \sigma_x(S)$ or at least it is as large as the number of elements in $R(v, X)$ that are not in the bigger union $U_{u \in TR} R(u, X)$. It follows that $\sigma_x(S \bigcup \{v\}) - \sigma_x(S) > \sigma_x(T \bigcup \{v\}) - \sigma_x(T)$, which is the defining inequality for submodularity.

Finally, we have $\sigma(A) = \sum_{outcomes X} Prob[X].\theta X(A)$. But a combination of non-negative linear submodular functions is also submodular. Thus the influence function for independent cascade model $\sigma(.)$ is submodular. Turning to NP-hard proof for the influence maximization problem under independent cascade model, this proof has been derived by considering the set cover problem that is defined by a collection of subsets $S_1, S_2, ..., S_m$ of a ground set $U = \{u_1, u_2, ..., u_n\}$ and we seek $k < n < m$ of the subsets that their union is equal to U. It is known that for an arbitrary Set Cover problem, a directed bipartite graph with $n+m$ nodes, where a node i corresponding to each set S_i, a node j corresponding to each element u_j, a directed edge (i, j) with activation probability $p_{i,j} = 1$ whenever $u_j \in S_i$. This problem is equivalent to deciding if there is a set A of k nodes in this graph with $\sigma(A) \geq n + k$. Remembering that the activation is a deterministic process. Thus, all probabilities are either

0 or 1. Initially activating the k nodes directly correspond to the set cover solution resulted in activating all n nodes corresponding to the ground set U. Therefore, any set A of k nodes has $\sigma(A) = n + k$, will solve the set cover problem.

[2] has introduced the first model and algorithm for learning influence maximization problem's parameters. The used technique to study the probability of models, which estimates users' action will be influenced by the neighbors, is by analyzing the actions' log of the social network and the network relationships or the network structure. For instance, in Figure 2.2 the undirected social graph contains 3 nodes and 3 edges with their time stamps to represent the time of creating the social tie; (b) a given action Log; (c) the constructed propagation graph for the action $a1$; (d) the constructed propagation graph for the action $a2$; (e) the constructed propagation graph for the action $a3$; and (f) is the Influence Matrix.

Furthermore, [2] has been developing static and time-dependent models for capturing influence, presented algorithms for learning the parameters of the various models and for testing the models. Nevertheless, a continues time model which achieved better accuracy but was expensive in contrast to the discrete time model which was the efficient model to test. Besides that, a testing algorithm has been developed and rarefied using Flicker dataset. A notable feature in [2] algorithms are they were optimized to minimize the scans over the action

log which is the key input to the problem of calculating the probabilities of influence. In fact, this is a significant feature since the action log tends to be extremely huge.

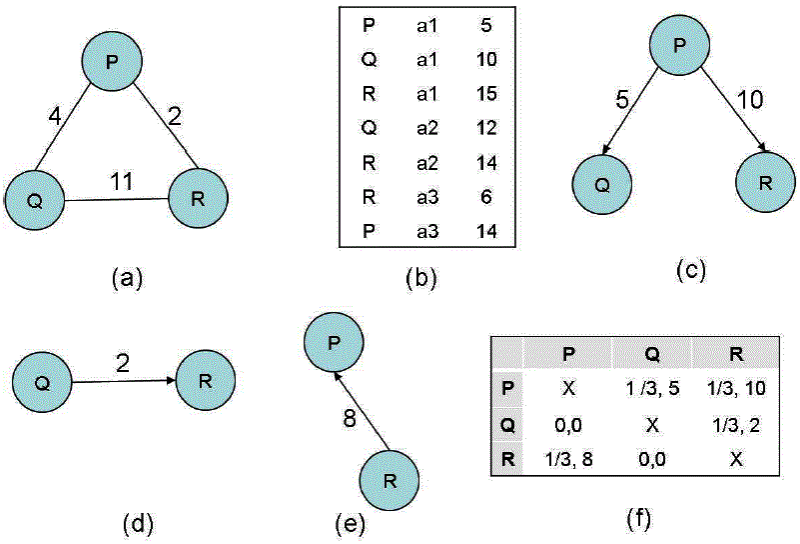

Fig. 2.2: Learning Influence Probabilities in Social Networks

Chen [32] proposed efficient algorithms and heuristics for the influence maximization problem by propose using degree discount heuristics that improves influence spread centrality-based heuristics nevertheless, when tuned for a specific influence cascade model, it accomplishes almost similar influence thread that obtained by the greedy algorithm. Not only that, but also the used algorithms suffer from a tradeoff in running time i.e. the first round is as slow the rest is much faster as a result of exploring smaller graphs. Before [32], Leskovec in "Cost-Effective Lazy Forward" (CELF) scheme, achieved a very

impressive result in speeding up by 700 times the selection of seed vertices process. This enhancement is resulted of utilizing the submodularity property of the influence maximization objective to rapidly reduce the number of evaluations on the influence spread of vertices. To improve this, a query approach for influence maximization [14] criticized the weakness in scalability of influence maximization algorithms is resulted in the lack in distinguishing specific users in social networks. Standing on the fact that states the importance of some users for various reasons. Thus, it is a better strategy to focus on maximizing the influence on the specific users. In [14], the influence maximization problem has been formulated as query processing to distinguish specific users from others. This distinguish been named target-aware viral marketing which includes people classifications to three categories. The first one is target users who have an interest in the item. While the second one is nontarget users who can be influenced for the item to introduce it to their friends. The last category is non-target users who are immune to being influenced for the item, because they do not want to introduce it to their friends. However, this classification deficient in neither practical explanation nor scientific analysis method.

To improve these results, many research have been done to analyse the structure of social relationships. It is worthy to spot the work has been done by [33] and [4]. In [33], the paper extended the classical voter model to incorporate negative relationships as a result of the fact that the relationships may

contain negative ones which basically means foe, spite, or just distrust. Similarly in e-Bay website that allows the agents to label each other by trust or distrust tags. However, another fact has not been tackled in this work which states that any person may adopt a suggestion from a foe if the person trust the foe experience. [33] extended the voter model in signed digraph to handle two opposite opinions. Moreover, [33] provided an interpretation for a random walk under their voter model. Their contribution in influence maximization problem devolves on calculating influence contribution for the network users. Particularly, it is decisive to calculate for each user the instant influence in each time step because the seed set will be the users that maximize the instant influence contribution. The limitations in citefriend and foe revolves around convergence time of voter model on signed digraphs as it shows no deference between signed digraphs or unsigned digraphs. In addition to limiting the influence diffusion in signed networks under the voter model, and the difficulties to be exported to other models like the independent cascade model, and the linear threshold model.

Differently, in [4] confirmed homophily principle where people favor the decision of the people whom share similar interest and opposing the idea of their foes. Nevertheless, they targeted to minimize the propagation cost of negatively influenced users. The used method in citerelationship types is assigning cost of propagation to each social link due to various factors including interaction

frequently, propagation delay, and social tie. Their solutions stands on the observation that confirmed the possibility to find a short path between two nodes without influencing any node against an idea. [4] confuted the method to influence the destination positively while controlling the number of negatively influenced intermediate nodes using the dynamic programming to eliminate the nodes that maximize the negative influence in the network paths. Turning to [10] which clarified a lack in the research that tackle signed network in either using structural balance theory or social status theory. This lack is the sign of an edge when its neighbors do not share a common neighbor or a common friend. It is important to clarify the meaning of structural balance theory or social status theory. Structural balance theory is more popular which refers to homophily principle and follow the saying "the friend of my friend is my friend", while structural balance theory means the initiator of the edge has a higher status and the edge has a positive sign while the edge from the recipient has a negative sign to indicate that the recipient is viewed to have a lower status with the assumption that the edge in two directions exists. [10] classified the nodes in the signed graph to 16 different types of node considering the signs of their edges in these types. It is remarkable to mention that node type features can be applied in structural balance theory or social status theory, to predict the sign of edges that do not have common neighbors. This classification has been derived using Bayesian node features

based upon partially observed signed directed networks.

Moreover, a topic influence model was studied in [22], and [3] to add a topic-modeling perspective of social influence by introducing topic-aware propagation models and identifying influencers in topic-specific networks. In [22], the authors introduced a new influence propagation model instead of considering user-to-user influence to overcome the scalability as well as increase the efficiency and lastly to avoid the risk of overfitting where random errors are described in the mode in addition to the actual relationship. In [22], the model considers users' authoritativeness and interest in a specific topic and how an item is relevant to a specific topic. This approach outperformed tractional (IC) and (LT) models in real world cascade. Also, the main benefits that have been observed in (AIR) is achieving consistent gain over the previous approaches. In (AIR), the likelihood propagation has been calculated using expectation maximization method (EM) because the formulation of influence maximization problem is not traceable in closed form. Expectation maximization method (EM) is iterative method for finding maximum likelihood estimation of the models parameters. In each iteration, the algorithm alternates between performing an expectation step, which creates a function based on the current log-likelihood estimation values for the parameters, and a maximization step, which computes the maximum expected log-likelihood parameters. Eventually, these parameters are used to maximize the likelihood

of the statistical model [30]. However, the (AIR) model has some limitations in term of dealing with new category of items or the items that have not been categorized in addition to considering this model as a static one regard the treatment of the time. This work was followed by [3] which proposed a general search framework for finding topic-specific key influencers with various models. In contrast, we show in this work 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, [5] 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. In detail, this paper attempts to model "The Tipping Point" theory by Malcolm Gladwell [19] where three categorise of users have been highlighted because of their crucial effect of the cascade. Those categories are "fascinating" people or mavens, connectors, and salesmen. In [5], the approach was studying social networks and the cascading behavior to extract properties that would help create successful information cascades taking into account the static characters of dynamic cascade including degree centrality and distance centrality. The main difference in this paper from "The Tipping Point" theory components is the need to introduce a fourth actor who is the translator in order to close

bridges between close-knit groups. In this paper, a translators who represents "role" more than "position", [5] used cascading behavior to "determine the communities and translators of the network rather than depending solely on the structure of the network".

Mavens have been introduced in [5] to represent users who seek new knowledge. They cannot directly help but instead they help others and therefore share the knowledge they acquire with others considering the fact that hearing something from a maven is very likely to believe the correctness and importance of this piece of information. Based on [5], to detect these users in a graph, mavens node start a large number of cascades as they represent the original source of new information and have high influence on their neighbors. This assumption does not precisely match the business definition of mavens. Thereby, this work redefine mavens in graph based on the used definitions in business field as it will be displayed later.

In summary, it is worth mentioning that none of these works developed an information cascade model under marketing characteristics identifications.

2.3. Mavens in Business Field

The first illustration of the market maven concept was in 1987 by Feick and Price [16]. 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) [26], [16] and the modern web-based channels [29], [18].

The authors of [29] confirmed the notable absence of these studies in the understanding of market maven behavior in alternative communication channels. Even though finding new ways of recognizing mavens and targeting them is increasingly important in the twenty-first century. This rises many questions around the way to identify the extent of market maven behavior is retained across physical, and social networks channels. In another word, how the behavior of a maven can be identified in the virtual world and how it differs than the physical one. This research confirmed that many behaviors of the market maven are constant across the physical world and virtual world as well. They also summarize the personal characteristics of market mavens across social networks. Furthermore, mavens are well-thought-out to be different to other influencers such as opinion leaders, innovators, and early adopters by observing their activities which can be spotted in their general knowledge instead of focusing on a particular product category. This characteristic is an important distinction because it is in contrast to innovators and opinion leaders who have

limited interest in general product categories and their focus spot on specific products or specific categories. In addition to spanning multiple product categories feature, market mavens have other characteristics which include the pleasure of sharing information in order to reinforce their image in their society, individualism, affinity for technology accompanied with an intensity of media usage, and knowledge of other mavens. Even though there is a difference between the physical world and social networks can distress the extent of the individuals interaction with the medium. Besides that, mavens personality considers to focuses on "objective and observable traits" as well as "roles, attitudes, goals, and behavioral tendencies". In fact, most current observations including [29] hold that mavens' personality is relatively predictable and stable across time. However, it is important to mention that market maven behavior in social networks does not always match the physical world channels. In the physical world, the authors emphasized on some propensities of individuals with high market maven including the high likelihood of having knowledge of other market mavens where they assess each other in raising awareness of goods and services as well high communication rate between mavens in order to evaluate goods and services comparing to individuals with low market maven propensity. Yet, this characteristic of mavens is less likely to be detected on the web and in virtual worlds related to real-life. Another characteristic has been studied by [29] is mavens are usually more competitive with a higher

degree of individualism. This catechistic is reveled in social networks by the high affinity for technology to forward, share and learn through the available technology. It seems likely that they will have a greater affinity to trial and adopt new technology than other users. Lastly, [29] mentioned that mavens in real life usually have greater channel experience with high intensity of usage of both the web and social networks than individuals with low maven propensity. Their finding for this characteristic is quite generic where mavens are not significant. Thus they suggested that use channel experience and use intensity did not significantly boost users to become mavens in the Web.

In [19], the authors underline the main motivations to perform research on mavens as their importance in many aspects as their importance of initiating discussion in shopping considering that they are a cheaper way to gain more customers. With SNSs users get benefits of the ease of communication and sharing information. However, targeting mavens with communications could be difficult if we overlook the mavens motivations which are their high sense of obligation to share information, their desire to help others, and their pleasure to share information. [19] confirmed that mavens are having similar behaviors that involve the communication about a product more than other, high media consumptions, the positive attitude toward advertising, and awareness about prices. In [18] the authors focus on how individualism contributes to explain mavens in word of mouth applications as a factor of individual differences. In

specific, the authors investigate the behavior to transmit word of mouth differ between mavens and other users. This study provides knowledge related to a type of social networks user who has a distinct psychological motive underlying their activity. [18] mentioned that mavens role in SNSs is based on their need for self-enhancement. Thus, when they need to spread the word about new products, mavens could be better off targeting them as opposed to the more elusive innovators or opinion leaders.

In [26], the authors clarify the importance to distinguish mavens of other critical uses like early adopters (innovators) and opinion leaders. For early adopters or innovators, the studies have found that early adopters have a tendency to be younger ages, better educated, and come from a higher social status relative to others in the society, who easily can afford to purchase multiple things. [26] also referred that early adopters are more likely to share their experiences about the products they gained with only their local reference groups. While opinion leaders difference in the characteristics that relay on levels of sociability and, such as to which extend they are open to share their personal experiences beside their degree of annotativeness, and the positive attitudes toward trying new things.

On the other hand, mavens differ from the previous two categories in many ways. One major different based on [26] is the higher levels of general knowledge in mavens about the marketplace and product marketing mixes like prod-

uct, prices, distribution, and promotions. Also, market mavens collect and exchange information about a wide spectrum of issues such as product quality, availability, and store personnel characteristics with other features that may be related to themselves and to other consumers. Based on this contrasts between mavens and the opinion leaders and early adopters who have deeper knowledge and want to share information about a specific range of products within specific people in their environment. 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. [31] 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, [29] 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 awareness about prices, as well as the pleasure in sharing information. In fact, [29] 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, Mavens are critical users in social networks that share many common characteristics; and the current research lacks utilizing marketing identifications to enhance the selection of seeds.

2.4. Modelling Psychological Attributes

In order to model psychological attribute, we need to identify psychophysical scaling which is identified as the physical correlate. In [28] fined the differences in psychological scaling should be equal psychological units, and they suggested to use a logarithmic relation between the psychological magnitude and the physical magnitude of a stimulus. To use psychological measurement some requirements has to be ensured. This requirement revolved around the stimulus of the physical correlate measurement. Also, the presentation of stimuli is relatively discriminable. Lastly, the assumption that that noticed differences are equal. A criticism of Fechner's rationale is that any functional dependence of a psychological unit of measurement on measuring the physical attribute is at the commencement including the solution to one of the problems we seek to study.

Another model has been suggested by Thurstone, in the Law of Comparative Judgment [17] that approach to use the standard deviation of the normal distribution to be the unit of psychological measurement. To apply that, two aspects should be considered in every type of scale. The first one is defining the formal system, which consists of elements, operations, and properties. While the second aspect to identify real objects systems with operations that can be performed on them and with properties that are experimentally observable. These two aspects will achieve the target by mapping the object system into an abstract system. To utilize this approach, it is necessary to differentiate between measurement and scaling. Measurement basically refers to the assignment of numbers to objects. Whereas scaling merely identifies classes of objects, taking into the account the purpose of this model is mapping between the numbers and equivalence classes of objects. If we succeed to combine measurement to the relation like greater than between these numbers, that means we introduce an ordinal scale. In the abstract system, operations of addition and multiplication on the numbers are permitted correspond to operations on the object system. Another scale can be used which is easier and most common which is the interval scale. In the interval scale, the differences between numbers are assigned to objects rather than on the numbers themselves. It refers to scales that fits some axiom systems but do not possess a constant and universal unit of measurement. The operations in this approach are exper-

imental, which means the objects establish an ordinal scale, in contrast, the distances between objects are at least partially ordered. The characteristics of using interval scale can be summarized in the following, using percentage to discriminate for the stimuli, any obtained data do not usually constitute a set of different scales which unfold into another possible scale.

Likewise, it is suitable to introduce the difference between the unidimensional case and the multidimensional case. In unidimensional case the first scale can be thought as picking up a continuum like finding the length of a string at the position of the individual accompany with giving the rank order of the stimuli in terms of their relative distance away from the individual in either direction. In the multidimensional case, the multidimensional surface which is being selected at the locus of the individual, and the scale gives the rank order of the stimuli in terms of their relative distance rom the individual in either direction. Another useful mode which is introduced by [6] is the summative model. This model is used in the partially ordered scale. In this model, the original components of a given type of behavior are assumed to compensate for each other, or be combined additively to construct the observed behavior. Focusing more on analyzing the behavior, [6] highlights a conjunctive behavior in which the behavior has several required attributes for each to a minimum degree. In the summative model, the lack of either primitive attribute means failure, and surplus of one does not compensate for a lack of the other. As

a result, it would be desirable to have an index to measure the concept of correlation between two the attribute and the scale. In this research, a summative model is used as mavens' behavior is a conjunctive behavior that will be explained later.

After all, [6] pointed out the dilemma of the social scientist that appears in the difference between any two used approaches will lead to a difference in the degree to which the integrity of the data is maintained. In addition to distinguishing the errors in the data.

3. PROBLEM DEFINITION

After highlighting the background about the influence maximization problem in social networks as well as emphasizing the importance of modeling psychological attributes of particular users called mavens; we will formulate this problem as follows. 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 tagging people, pages, or adding locations. For instance, table 3.1 illustrates the basic actions in the Facebook [34]. In specific, some actions have benefits in joining couple of users together like the action "og.follows" as it refers to an action that a user follows another Facebook user.

Name	Description
books.quotes	An action representing someone quoting from a book.
books.reads	An action representing someone reading a book.
books.wants_to_read	An action representing someone wanting to read a book.
fitness.bikes	An action representing someone cycling a course.
fitness.runs	An action representing someone running a course.
fitness.walks	An action representing someone walking a course.
games.achieves	An action representing someone reaching a game achievement.
games.celebrate	An action representing someone celebrating a victory in a game.
music.listens	An action representing someone listening to a song, album, ect.
music.playlists	An action representing someone creating a playlist.
news.publishes	An action representing someone publishing a news article.
og.follows	An action representing someone following a Facebook user.
og.likes	An action representing someone liking any object.
pages.saves	An action representing someone saving a place.
restaurant.visited	An action representing someone visiting a restaurant.
restaurant.wants_to_visit	An action representing someone wanting to visit a restaurant.
sellers.rates	An action representing a commerce seller has been given a rating.
video.wants_to_watch	An action representing someone wanting to watch video content.
video.watches	An action representing someone watching video content.

Table 3.1: Actin Types in Facebook Open Graph

Also, each action has the format $a_i = [a_{i,property(1)}, .., a_{i,property(n)}]$. Table 3.2 demonstrates an example of a possible action vector components in Facebook Open Graph [34].

Name	Type	Description
restaurant	Reference	The restaurant that was visited.
created_time	DateTime	The time that the action was created.
end_time	DateTime	The time that the user ended.
expires_in	Integer	The expire time of the action from the $publish - time$.
$fb : explicitly\_shared$	Boolean	The user is explicitly sharing this action.
message	String	A message attached to this action.
no_feed_story	Boolean	Do not post this action to the feed.
place	Place	The place that the action took place.
ref	String	A 50 character string identifier for tracking and insights.
start_time	DateTime	The time that the user started.
tags	Array¡Profile¿	Any other users that performed the action.

Table 3.2: An Example of a Facebook Action Vector

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 3.1. *(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: $A_u^* = \{a_i | \forall v \in N(u), if \exists (u, a_i, t_u), (v, a_i, t_v) \in ActionLogthen, t_u < t_v\}$. 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: $A_u^- = \{a_i | \forall v \in N(u), (u, a_i, t_u) \in ActionLog, (v, a_i, t_v) \notin ActionLog\}$ The user individualism ind_u can be measured as follows:

$$ind_u = \frac{\mid A_u^* \cup A_u^- \mid}{\mid A_u \mid} \tag{1}$$

Definition 3.2. *(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. $\hat{A}_u$ is expressed as follows: $\hat{A}_{u,v} = \{a_i | \exists v \in N(u), (u, a_i, t_u) \in ActionLog : a_i = [a_{i,property(1)}, .., a_{i,property(n)} = v]\}$. The users sharing desire shr_u can be measured as follows:

$$shr_u = \frac{\mid \sum_{v \in N(u)} \hat{A}_{u,v} \mid}{\mid A_u \mid} \tag{2}$$

Definition 3.3. *(User's multiplicity of spanning)* Let $Topic$ represent all topic categories in the given action log and let $Topic_u$ 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:

$$mul_u = \frac{\mid Topic_u \mid}{\mid Topic \mid} \tag{3}$$

After defining the mavens' main characteristics in social networks, a conjunctive behavior of the maven requires these three attributes (user's individualism, user's sharing desire, and user's multiplicity of spanning) for each user to a minimum degree. We create for each user u a vector of data $[ind_u, shr_u, mul_u]$

which represents numerical features for each user in the social network. This numerical representation of users is called a feature vector which facilitate processing of machine learning algorithms and statistical analysis. This feature vector summarizes the maveness confidence level for a user u. Users' characteristics lead to redefining the social graph based on the maveness confidence level.

Definition 3.4. *(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 $M : [ind_u, shr_u, mul_u] \rightarrow R$ 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:

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

where Δ 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 Δ.

The maven graph consists of users with their confidence 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 brief, multicommodity flow problem aims to find a feasible flow in a given directed network with edge capacities and a given set of commodities which is the message or the idea in this research, where a commodity is significantly affected by the contact frequency and considering the constraints about the edges capacities. 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 in the network, where generally assume that information is flowing through the edges and 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 $M : [ind_u, shr_u, mul_u] \rightarrow R$ 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 Ω 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:

$$max \sum_{v \in Mav} f(v,u), s.t. \sum_{v \in Mav} f(v,u) \leq m_v, f(v,u) \geq 0. \quad (5)$$

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.

4. 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

$$\begin{aligned}\Pr(v_{active}|u_{active}) &= \frac{\Pr(v_{active}).\Pr(u_{active}|v_{active})}{\Pr(u_{active})} \\ &= \frac{\Pr(u_{active} \cap v_{active})}{\Pr(v_{active}).\Pr(u_{active}|v_{active}) + \Pr(v_{inactive}).\Pr(u_{active}|v_{inactive})} \\ &= m_v m_u d_{v,u}\end{aligned} \tag{6}$$

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 $U = \{U_{active}, U_{inactive}\}$ to set multiple neighbors U spreading information to a user

v independently of each other. Thus, the joint probability $\Pr(v_{active}|U_{active})$ will be calculated as follows:

$$\Pr(v_{active}|U_{active}) = \sum_{i=1}^{|U|} \Pr(v_{active}|U, U_{active}).\Pr(U|U_{active}) = \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{m_v.d_{v,u}}{d_{u,v}} \tag{7}$$

Subsequently, we have to ensure that the proposed model satisfies the occurrence of a fully revealing informational cascade with any given priors [13]. Thus, we introduce the following theorem.

THEOREM 4.1. *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 $\Pr(v_{active})$ 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 [8], we will get

$$\Pr(v_{inactive}) \leq \Pr(v_{active}) \tag{8}$$

using Lemma 2 in [13] and inequality 8, we will define ε' and ε'' as follows:

$$m_v - \varepsilon' \leq \Pr(v_{inactive}) \leq \Pr(v_{active}) \leq m_v + \varepsilon'' \tag{9}$$

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:

$$m_v - \varepsilon' \leq \frac{\sum_{u=1}^{U} d_{v,u}.(1 - m_u)}{|U|.\sum_{u=1}^{U} d_{u,v}} \leq \frac{\sum_{u=1}^{U} d_{v,u}.m_u}{|U|.\sum_{u=1}^{U} d_{u,v}} \leq m_v + \varepsilon'' \tag{10}$$

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

$$d_{u,v} \geq 0 \forall (u,v) \in E \sum_{u=1}^{|U|} d_{v,u}(1 - m_u)(-m_v + \varepsilon') \geq 0 \sum_{u=1}^{|U|} d_{v,u} m_u(-m_v - \varepsilon'') \leq 0 \tag{11}$$

By applying Frkas lemma [27] to solve these linear inequalities, we will use κ_u for all $u \in U$ as a dual variable for $d_{u,v}$ to obtain the following equations:

$$\kappa_u + (1 - m_u)(-m_v + \varepsilon')\kappa = d_{v,u} m_u(-m_v - \varepsilon'') \forall (u,v) \in E \tag{12}$$

Since $\frac{m_u(-m_v+\varepsilon')}{(1-m_u)(-m_v-\varepsilon'')}$ is increasing in m_u, there is a κ that satisfies the system, where

$$\frac{\Pr(v_{active})(\Pr(v_{active}) - m_v - \varepsilon')}{\Pr(v_{inactive})(\Pr(v_{active}) - m_v + \varepsilon'')} \leq \frac{\Pr(v_{inactive})(\Pr(v_{inactive}) - m_v - \varepsilon')}{\Pr(v_{active})(\Pr(v_{inactive}) - m_v + \varepsilon'')} \tag{13}$$

which satisfies the necessary and sufficient condition for a fully revealing informational cascade based on theorem 2 in [13]. □

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

$$f_{(u,v)} = \frac{m_u.m_v.d_{v,u}}{d_{u,v}} \tag{14}$$

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 [24] 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.

4.1. Learning Personalized Mavens Characteristics from Action Logs

As a result of the users' characteristics measurements, we can create for each user u a vector of data $[ind_u, shr_u, div_u]$ 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 Figure 4.1, which makes it suitable for a class model in the feature space [30]. Basically, a Gaussian mixture model is a weighted sum of given feature components that have Gaussian densities, as given by the equation

$$E(x|\theta) = w.g(x|\mu, \sigma) \tag{15}$$

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 μ is the mean value and σ is the

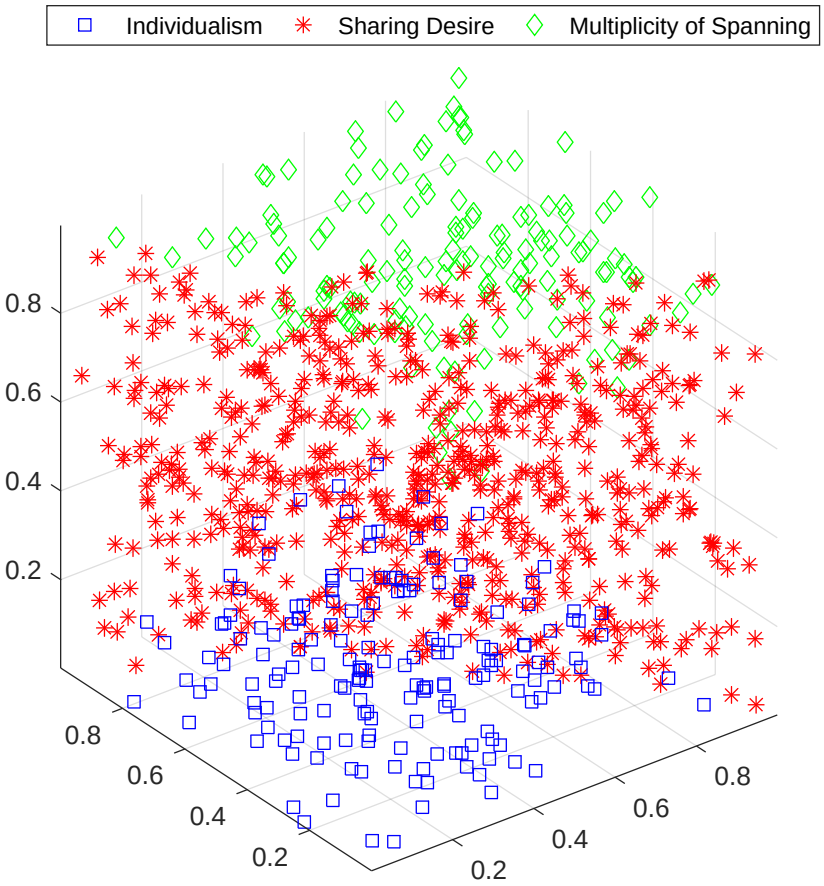

Fig. 4.1: Mavens Feature Space

covariance matrix. Each component density can be obtained by the following equation,

$$g(x|\mu,\sigma) = \frac{1}{(2\pi)^{3/2}|\sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)'\sigma^{-1}(x-\mu)\right\} \tag{16}$$

considering θ includes the values of w, μ ,and σ.

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 [30] to calculate the likelihood function:

$$\ell(X|\theta) = \prod_{n=1}^{N} E(x_n, \theta) \tag{17}$$

where X is a set of independent samples $X = \{\mathrm{x}_1, ..., \mathrm{x}_N\}$ used by a probability density function and the objective is to find θ that maximizes the likelihood:

$$\theta_{opt} = \max \ell(X, \theta) \tag{18}$$

The purpose of using the expectation maximization (EM) algorithm is that the above function is a non-linear function of the parameters θ. Thus, it is not possible to obtain a direct maximization. In the EM algorithm we will initially begin with a random guess for θ, which leads to get a new value $\overline{\theta}$, that satisfies $\ell(X|\ \bar{\theta}) > \ \ell(X|\theta)$.

The iteration will be repeated until an acceptable threshold value of convergence is reached. Eventually, the mixture weights will be:

$$\bar{w}_i = \frac{1}{N} \sum_{n=1}^{N} p(i|x_n, \theta) \tag{19}$$

After performing (EM) algorithm in the users dataset, we get discriminated nodes to mavens and ordinary nodes as shown in Figure 4.2.

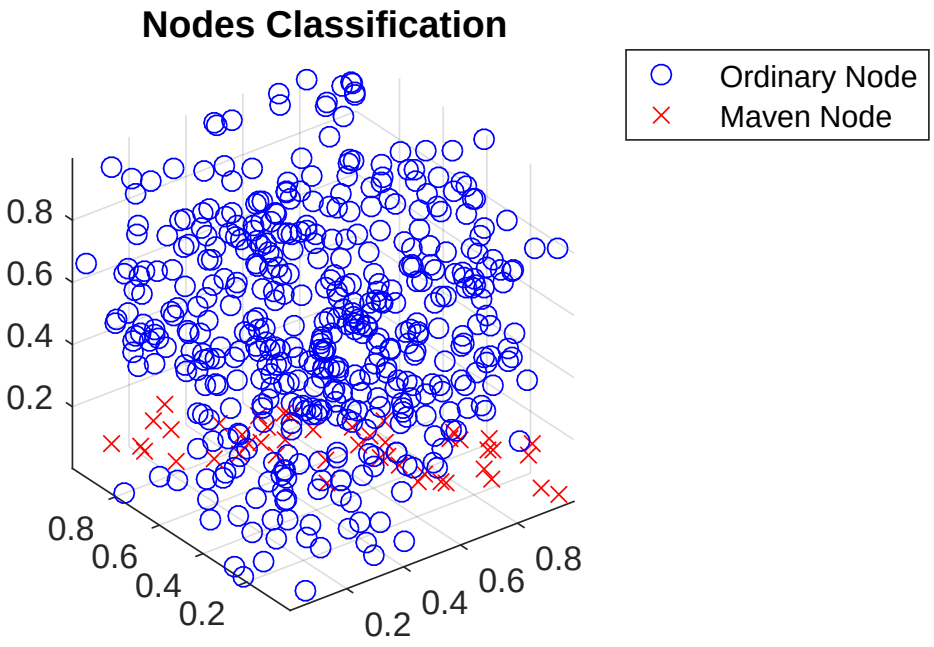

Fig. 4.2: Nodes Classification using Gaussian Mixture Model

5. 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 15, where θ_i is the previous estimate of the distribution parameters, and θ is a variable for a new estimate that describes the (full) distribution. Precisely, ℓ in equation 17 calculates the likelihood of the data, considering the unknown class with respect to the current estimate of the distribution described by θ_i. The M-step (lines 8-17) is used to maximize $Q(\theta;\theta_i)$ with respect to equation 18. 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 θ (line 18). The time complexity of Algorithm 1 is $O(3|V| + 2|V|^2)$.

Algorithm 2 creates the mavens 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 length. Thus, Algorithm 2 creates a path with a

Algorithm 1: Calculating Maveness Confidence Level of Nodes

Input: $u_{input} = [[ind_u, shr_u, mul_u]]$ feature vector of users with size $|V| \times 1$; ϵ : suitable convergence criterion

Output: $u_{output} = [[m_u]]$ maveness confidence level of u of users with size $|V| \times 1$

initialization: assign Θ_i, α_i to any value

repeat

  $Q(\Theta, \Theta_i) = 0$

  for $u \in u_{input}$ **do**

    $w_u = {\alpha_i}^{maven} p(u|maven; \Theta_i)$

    $Q(\Theta, \Theta_i) = Q(\Theta, \Theta_i) + w_u.(ln(w_u) + ln(1 - w_u))$

  end

  $\alpha_{i+1} = 0$ $\mu_{i+1} = 0$ $\sigma_{i+1} = 0$

  for $u \in u_{input}$ **do**

    $\alpha_{i+1} = \alpha_{i+1} + w_u$

    $\mu_{i+1} = \mu_{i+1} + u.w_u$

    $\sigma_{i+1} = \sigma_{i+1} + w_u(u - \mu_{i+1})(u - \mu_{i+1})^T$

  end

  for $u \in u_{input}$ **do**

    $\mu_{i+1} = \frac{mu_{i+1}}{\alpha_{i+1}}$

    $\alpha_{i+1} = \frac{\alpha_{i+1}}{|V|}$

    $\sigma_{i+1} = |V|.\sigma_{i+1}.\alpha_{i+1}$

  end

until $\Theta_{i+1} - \Theta_i > \epsilon$

$m_u = w_u \frac{1}{\sqrt{(2\pi)|\sigma|}} \exp\left\{-\frac{1}{2}(x - \mu)\sigma^{-1}(x - \mu)\right\}$ Return $[[m_u]]$;

positive length between mavens and the remaining nodes.The time complexity of Algorithm 2 is $O(|V|^2)$.

Algorithm 2: Creating Mavens Graph

Input: G=(V,E,M,D),Ω :maveness confidence threshold value

Output: $\grave{G} = (\grave{V}, \grave{E}), P$

initialization: $\grave{V} = \emptyset$; $\grave{E} = \emptyset$; $P = \emptyset$

```
1  for v ∈ V do
2      if m_v ≥ Ω then
3          V' = V' ∪ {v} ∪ {u|u ∈ N(v)}
4      end
5      for v' ∈ V' do
6          l(v,v')=d_{v,v'} / d_{v',v}
7          if l(v, v') > 0 then
8              P = P ∪ {(p(v, v'), l(v, v'))}
9              { E' = E' ∪ {(v, v')} ;
10         end
11     end
12 end
13 Return V', E', P;
```

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 add the value of $(1 + \frac{m}{m_u l(u,v)})$ to the current distance value (line 7). The procedure stops after processing all of the paths or when the smallest distance between all of the nodes is less than 1 (line3). The time complexity of Algorithm 3 is $|M||V||E|log(L)T_{sp}$, where M is the mavens number, V is the node number, E is the edge number in the mavens graph, L is the maximum number of edges on any mavens-terminal node path, and T_{sp} is the time required to compute the shortest s-t path in a graph with non-negative edge-weights.

Ultimately, Algorithm 4 identifies the k-node set based on the mavens graph. The k-node set is selected as follows. First, we mark V_{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 3: Finding Maximum Multi-commodity Flow in Mavens Graph

Input: G=(V',E'),P

Output: max flow f, min length l

initialization: $dist(u,v)$=min $l(u,v)$; $f = 0$

for $p(u,v) \in P$ **do**
  $p_{selected} \leftarrow$min $dist(u,v)$
  while $dist(u,v) < 1$ **do**
    $m \leftarrow$min m_i in $p_{selected}$
    $f = f + m$
    for (u,v) *in* $p_{selected}$ **do**
      $dist(u,v) = dist(u,v) + (1 + \frac{m}{m_u l(u,v)})$
    end
    $p_{selected} \leftarrow$ min $dist(u,v)$
  end
end
Return f, $dist$;

6. EXPERIMENTS

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

Algorithm 4: Identifying the k-node Set

Input: V',k

Output: S

initialization: $V_{temp} = V'$

while $|S| < kandV' \neq \emptyset$ **do**

  $u = max(f(u \in V' \cap V_{temp})$

  $V_{temp} = V_{temp}|\{v|v \in p(u,x)\}$

  $S = S \cup \{u\}$

end

Return S ;

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.

6.1. 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[1]. This

[1]www.kddcup2012.org/c/kddcup2012-track1/data

dataset includes more than 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 timestamps.

Giving more detail about the components of dataset and the way of constructing the needed information of this dataset. First, the action log which is described in table 6.1, noticing that the result field could have 0 value to represent that there is no response from the receiver user to what has been initiated by the sender user. Otherwise, it could be 1 to represent the positive response or -1 to indicate the rejection by the receiver user. The user action table has been divided into two tables one of them has been used in the training data set for nodes classification while the other one has been used to test and verified the classification.

UserId	ItemId	Result	*timestamp*

Table 6.1: Tencent Weibo Users' Action log

In addition, a deeper information about users is in the user profile data file as shown in table 6.2 where each user has a record contains the year of birth, the gender, the number of tweets which is a critical field to represent the total

actions that have been performed by that user, and the tag_Ids which can be used in the calculations of the number of topics that are discussed by that user. For instance, if a user likes hiking and swimming, the user may select "forest hiking" or "swimming" to be a tag. Indeed, there are some users who did not anything thus Tag_Ids will be 0. Each tag has been encoded in the dataset to a unique integer. Tag_Ids will be listed in the form "$tag-id_1; tag-id_2; ...; tag-id_N$". Another topics data are located in $user\_key\_word$ file. In this file, there are keywords that have been extracted from the textual content of the user (i.e. tweet, retweet, or comment). In the dataset the Keywords field has the following format "$kw_1 : weight_1; kw_2 : weight_2; kw_n : weight_n$". These weights give a precise representation of user usage. However, the weight attributes have been ignored in the experiment because our definitions do not emphasize the intensity of the usage of any topic.

UserId	Year_of_birth	Gender	$Number\_of\_tweet$	Tag_Ids

Table 6.2: Tencent Weibo Users Profile

UserId	Keywords

Table 6.3: Tencent Weibo Keywords

Beside users profile, items in the dataset have been gathered in the Item table as illustrated in table 6.4 to keep an information of each item in term

of its category and the used keywords. In the experiment, we paid a special attention of $Item-Keyword$ field because those keywords are extracted from the corresponding users profile by matching them with $Item-Category$ which is accounted in the users topics calculations. The format of $Item-Keyword$ is "$id_1; id_2; ...; id_N$", to represent each keyword in a unique integer value.

ItemId	$Item-Category$	$Item-Keyword$

Table 6.4: Tencent Weibo Item table

On the other hand, User action file as shown in table 6.5 has summarized statistics about the number of shared actions between each two connected users which highly critical in sharing desire characteristics calculation.

UserId	$Action-Destination-UserId$	$Number-of-action$

Table 6.5: Tencent Weibo Action File

Lastly, social network file is described in table 6.6 that constructs the directed graph from the users' follow history considering that the following relationship can be reciprocal.

$Follower-userid$	$Followee-userid$

Table 6.6: Tencent Weibo Social Network File

We conducted the experiments on an Intel(R)Core(TM)i7-4510U 2.6 GHz CPU machine with 8 GB RAM.

6.2. Nodes Classification and Maveness Confidence Level Calculation

The experiments start by calculating the mavens characteristics rates for the dataset users. Figure 6.1, 6.2, and 6.3 illustrate the density distribution of the defined characteristics of mavens. The first characteristic, the individualism, is shown in Figure 6.1. This figure reveals that the highest densely individualism ration is between 0.4 and .5. It is worthwhile to note that this characteristic clearly distinguish the users. Moreover, there are only 2.E+5 users with a high individualism rate.

Turning to Figure 6.2 that represents the distribution of the sharing desire ratio. In fact, the range between 0.6 and 0.7 has the highest users density. In general, users are evenly distributed in the sharing desire rates. It is also noticeable that the rate between 0.8 and 0.9 has the lowest density which was less than 1.5E+5 users.

Moreover, Figure 6.3 highlights the density distribution of users in term of the multiplicity of interest rates. In fact, there are two rates get the densest users. These rates are between 0.4 and 0.5 as well as 0.2 and 0.3. Therefore,

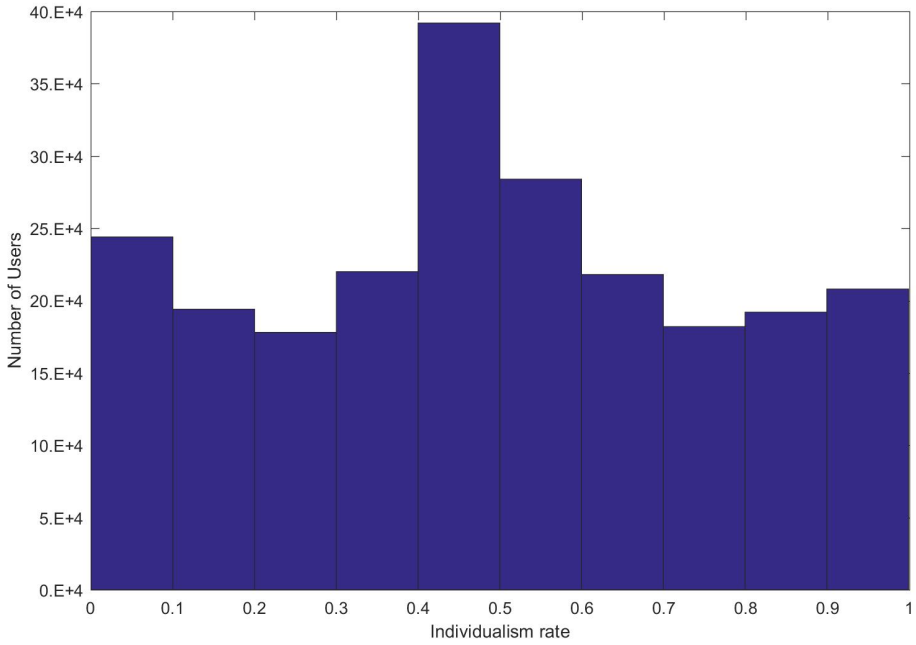

Fig. 6.1: Individualism Density Distribution

the multiplicity of interest is the least discriminative characteristic in term of users classification.

After calculating the feature vector value for each user, we classified the network users using Algorithm 1. Figure 6.4 demonstrates the ratio between mavens and the total network users in the given dataset. In fact, using a loose threshold value which is 0.65 leads to getting 160 mavens in total of 20000 network users. While with a bigger number of dataset users (around 400000) the classifies mavens reach 1958. In contrast with 0.9 threshold value, the number of mavens was only 375 within the same 400000 users which accounted

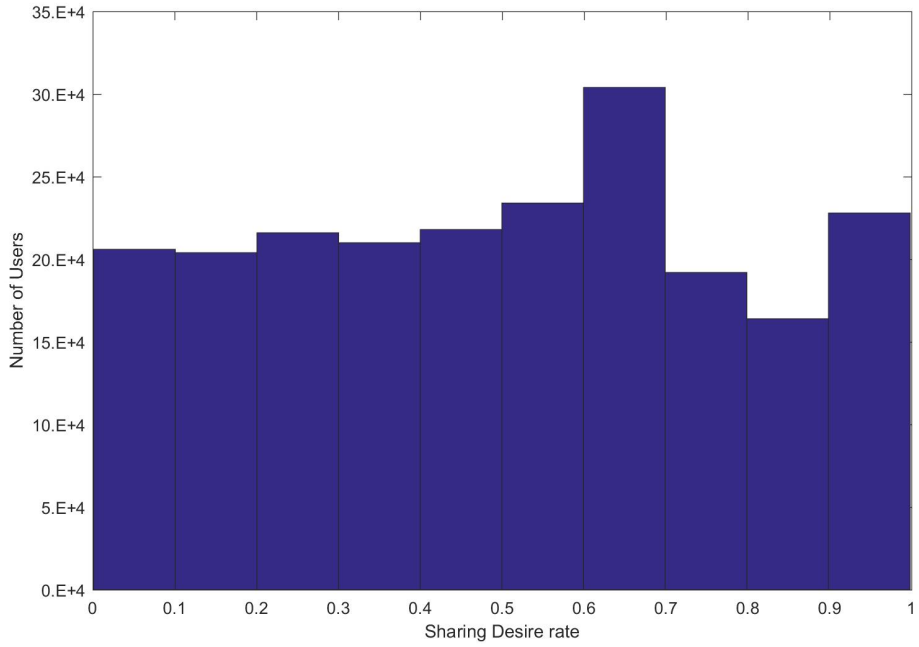

Fig. 6.2: Sharing Desire Density Distribution

as a narrow classification. Nevertheless the threshold value 0.8 achieved a similar number of mavens to 0.9 threshold with the smaller dataset which is around 60 mavens, it added more mavens with the larger dataset. It classified around 950 mavens in 400000 network users.

In addition, Figure 6.5 shows the correlation between the users activities and the maveness confidence level. It is remarkable that users with a high maveness level tend to have a high level of sharing activities which reach around 80% of their activities. However, users with low maveness confidence level have an average sharing level near to 20%. On the other hand, in term of

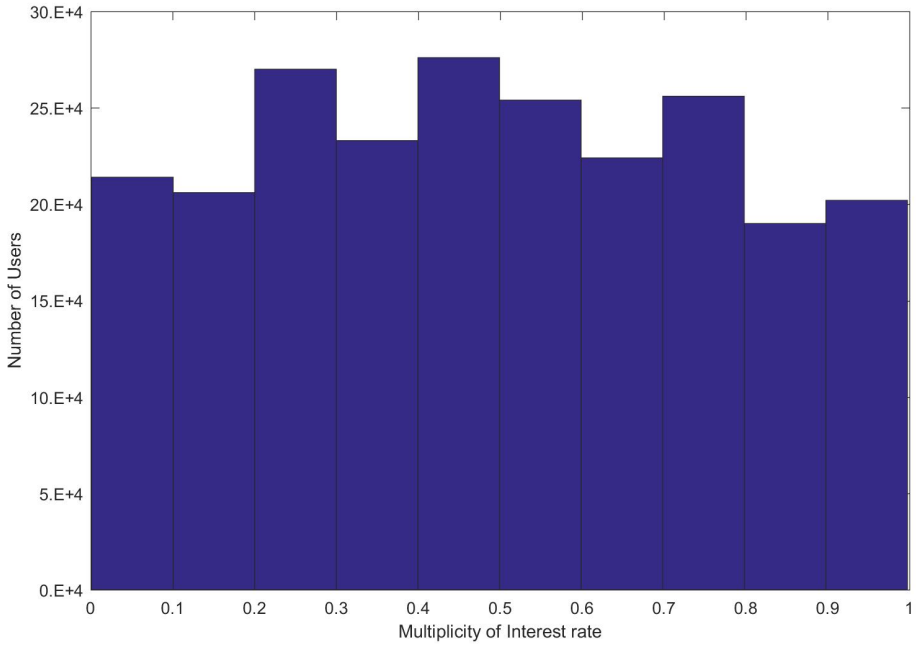

Fig. 6.3: Multiplicity of Spanning Density Distribution

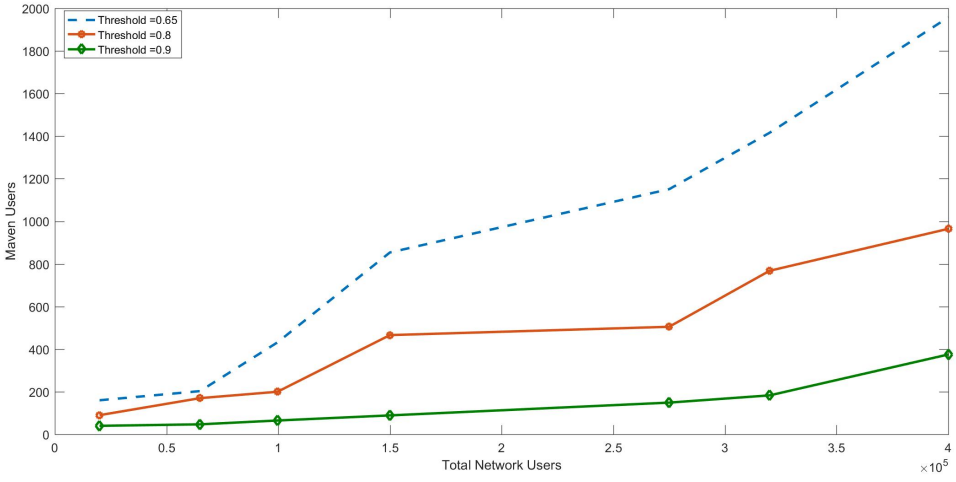

Fig. 6.4: Mavens Ratio Among Network Users

receiving information activities, users with high maveness level score around 20% which is compatible with their high level of individualism characteristic. It is also noticeable that users with low maveness confidence level has more than 80% of receiving information activities.

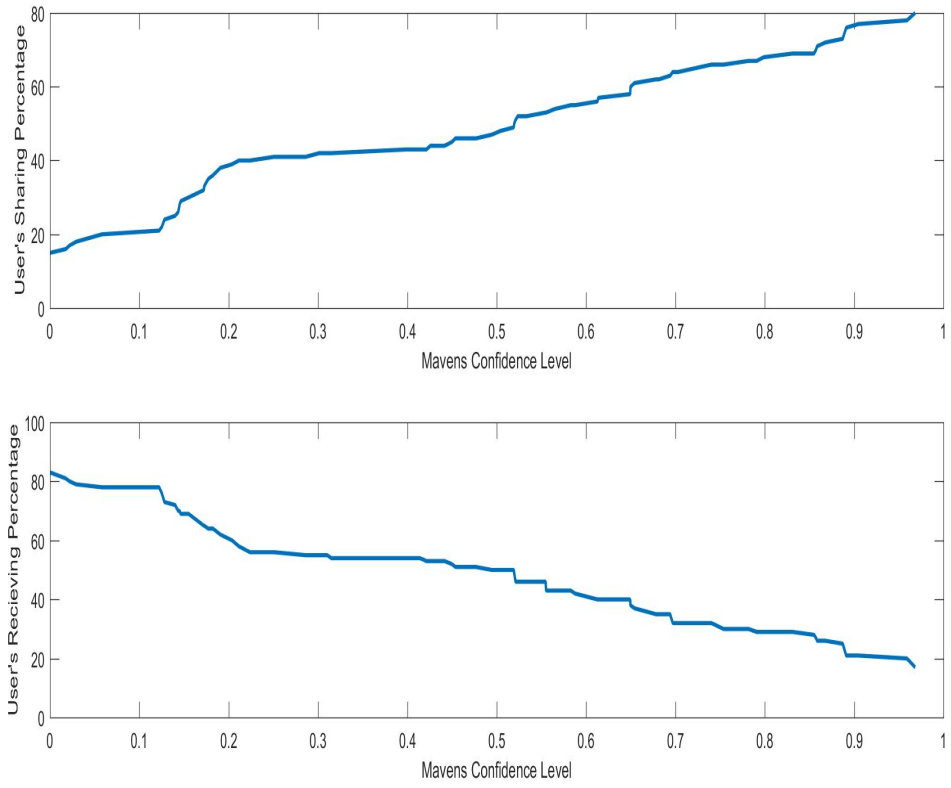

Fig. 6.5: Users Activities and Maveness Correlation

6.3. Accuracy of Learning

To evaluate the learning of users' characteristics considering that 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 mavens model in [5] 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 mavens definition in [5] to test the same amount of actions. Figure 6.6 illustrates that our definition for a maven performed better in estimating users' action behavior than the definition in [5]. Hence, the definition in [5] only highlights the top influencer users, and the test allows us to evaluate the contribution of mavens characteristics modeling to the prediction of users' actions.

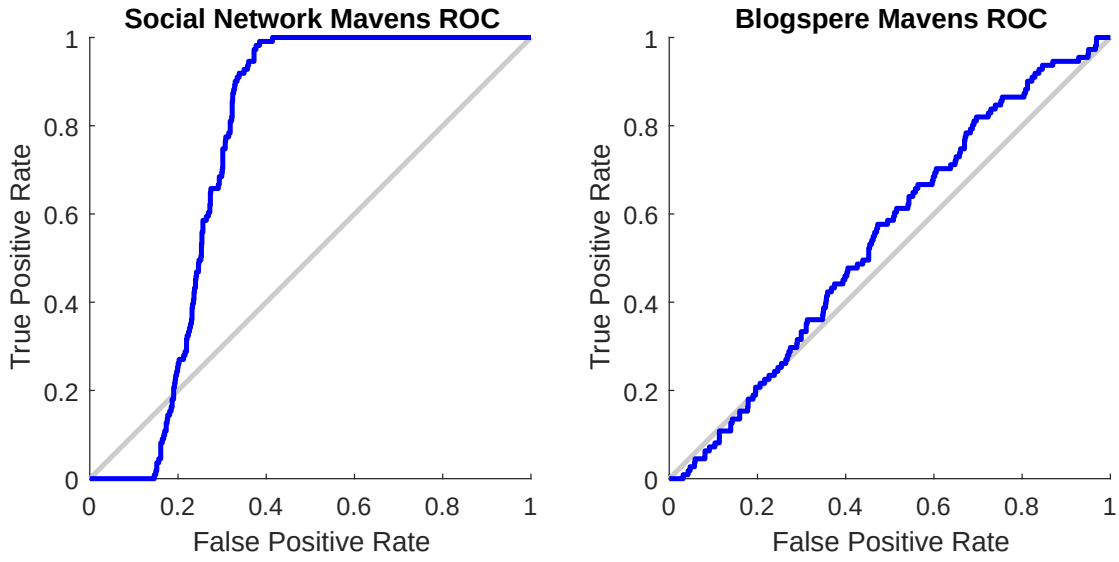

Fig. 6.6: ROC Analysis

6.4. 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

Ω threshold values. Figure 6.7 reveals the degree feature between the original social graph and the mavens graph in both outgoing and incoming edges. In the left figure, the maximum outgoing degree of a node reaches 350 in front of only 245 maximum degree of a node in the maven graph. Thus figure confirms that maveness level of any node is not reflected by its degree or its ability to reach the largest number of nodes. In addition, the original graph has a round 1.E+6 nodes who have only one connected neighbor. In contrast of maven graph that has less than 4000 nodes that have only one connected neighbor. As a matter of fact, the nodes with the smallest number of neighbors will not be the desirable nodes in the cascading process. On the other hand, the right figure which represents the incoming degree of nodes in the above mentioned graphs. It is clear that the incoming degree of the original graph is slightly lower than the outgoing degree. However, the incoming degree in the mavens graph is significantly lower than the outgoing degree. This low values in the incoming degrees resulted of the elimination of all edges that are not participated to create paths to mavens. As a result, the incoming degree figure precisely reflects the effectiveness of mavens graph.

As a consequence, we studied the difference between the length of the paths in the original social graph and the the mavens graph as result of its importance to evaluate the complexity of many calculations like the flow calculation

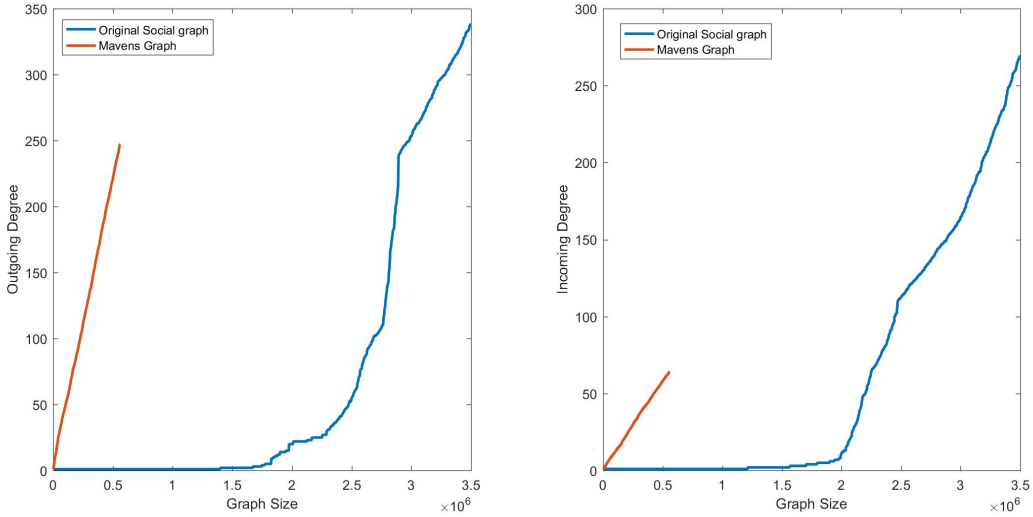

Fig. 6.7: Mavens Impact on Nodes Degrees

and finding the possible activated nodes or the marginal gain. Figure 6.8 reveals the difference between the length of paths in those graphs. In the original social graph, there are some nodes can create paths with length 800. This number emphasizes the high complexity to perform any estimation on the graph. However, the maximum path length in the mavens graph does not exceed 200. This number has a noticeable impact in reducing the computational cost of calculation information flow as well as finding the possible activated nodes. Beside that, around 12.E+5 nodes in the original graph have less than 3 path length which indicates that nodes are not effective in information cascade process. In contrast, half of the nodes in the mavens graph have a desirable length which exceeds 100 path length. Another remarkable feature is the big

difference between the participated nodes in the original social graph that have paths with 200 length while the quarter of this number in the mavens graph have path with the same length. The length of the path is consistent with mavens model as it targets the nodes that can stimulate the information cascade.

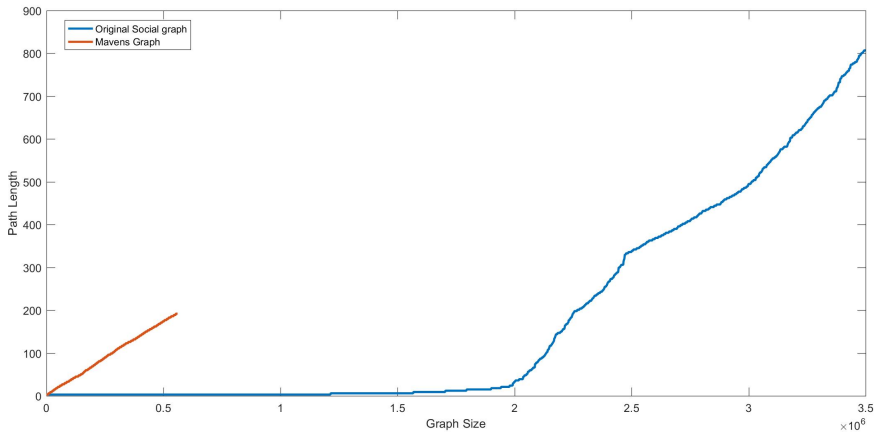

Fig. 6.8: Mavens Impact on the Paths Length

Meanwhile, as shown in Figure 6.9, 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 mavens graph size based on the seed size. Thus, this point will be considered in future work.

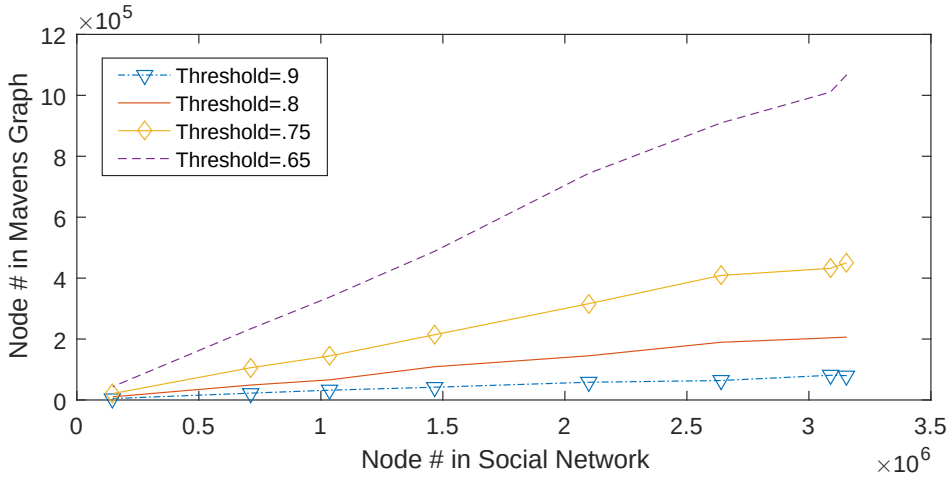

Fig. 6.9: Mavens Impact on Reducing the Size of Social Graph

6.5. 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 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 Figure 6.10, 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.

Additionally, Figure 6.11 clarifies the comparison between the number of successful attempts that are initiated by mavens in front of the number of blogs that are resulted of mavens cascade. In our model, top mavens accomplished around 140 successful attempts which slightly outperform blogsphere mavens achievement who were successfully generating around 100 blogs in a

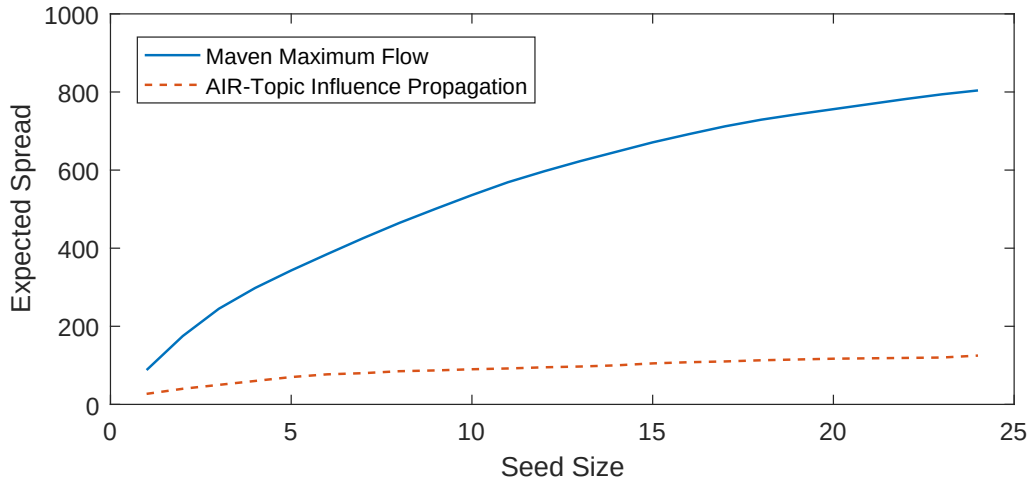

Fig. 6.10: Information Diffusion Experiments

single cascade [5]. Furthermore, the difference in the activations between these models steadily increases within the top 500 mavens. After that, our model maintain higher level of average positive activations in contrast to blogosphere mavens where the average cascaded blogs drop dramatically. In other words, for 1500 users in the maven graph, th average successful attempts of the users was 15 against just 3 cascaded blogs within the same number of mavens. To sum up, the decrease in the number of successful attempts in mavens model is generally similar to the average cascaded blogs in blogosphere model with the top 500 mavens.

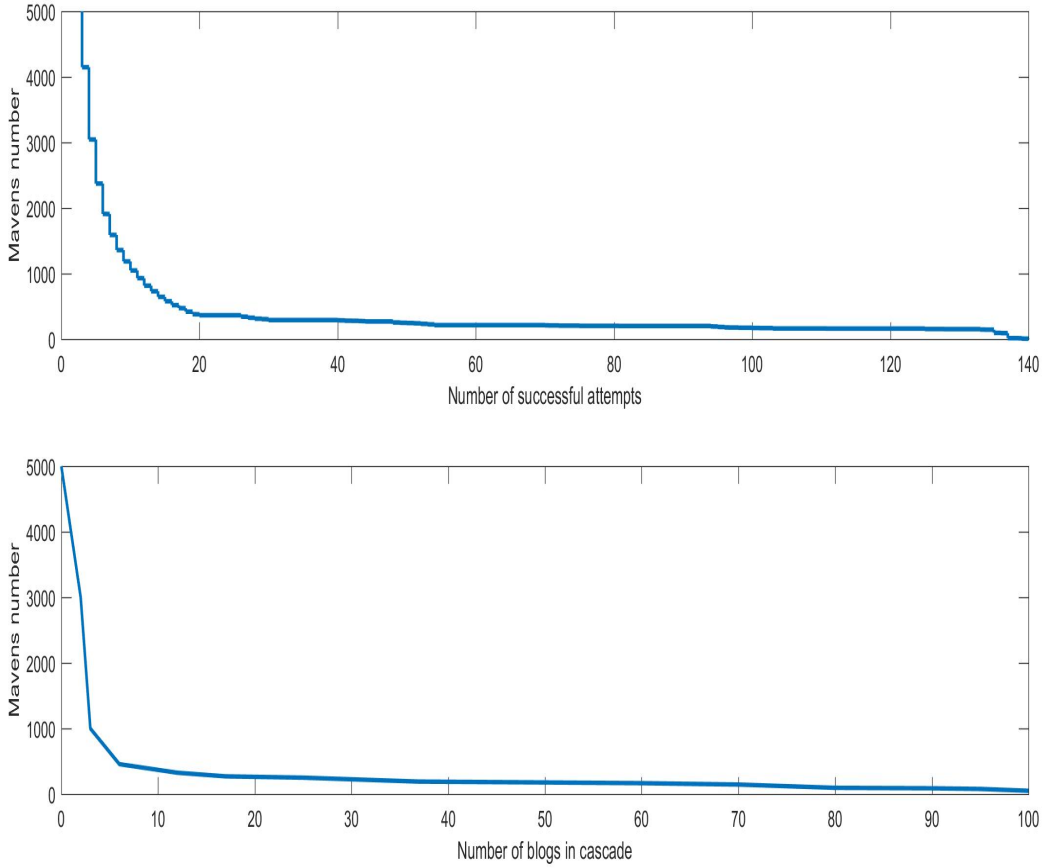

Fig. 6.11: Mavens Success Measurements

7. CONCLUSION

This research 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 reviled

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 thesis. 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

[1] A. Goyal, F. Bonchi, L. V. Lakshmanan, S. Venkatasubramanian, "Approximation analysis of influence spread in social networks", 2005.

[2] A. Goyal, F. Bonchi, and L.V. Lakshmanan, "Learning influence probabilities in social networks", in: Proceedings of the third ACM international conference on Web search and data mining, 2010, pp. 241-250.

[3] B. Bi, Y. Tian, Y. Sismanis, A. Balmin, and J. Cho, "Scalable topic-specific influence analysis on microblogs", 2014, pp. 513–522.

[4] B. Guler, B. Varan, K. Tutuncuoglu, M. Nafea, A. A. Zewail, A. Yener, and D. Octeau, "Communicating in a socially-aware network: Impact of relationship types", in Signal and Information Processing (GlobalSIP), IEEE Global Conference, 2014, pp. 788-792.

[5] C. Budak, D. Agrawal, and A. El Abbadi, "Where the blogs tip: connectors, mavens, salesmen and translators of the blogosphere", in Proceedings of the First Workshop on Social Media Analytics, 2010, pp. 106–114.

[6] C. H. Coombs, "Mathematical Models in Psychological Scaling", Journal of the American Statistical Association, vol.46, no. 256, Dec.1951, p. 480.

[7] C. McCarthy, "Facebook: one social graph to rule them all", 2010.

[8] D. Easley and J. Kleinberg, "Information cascades, Networks, Crowds, and Markets: Reasoning about a Highly Connected World", Cambridge University Press, 2010.

[9] D. Kempe, J. Kleinberg, V. Tardos, "Maximizing the spread of influence through a social network", in: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, Washington D.C., 2003, pp. 137–146.

[10] D. Song and D. A. Meyer, "A model of consistent node types in signed directed social networks", in: Advances in Social Networks Analysis and Mining (ASONAM), IEEE/ACM International Conference, 2014, pp.72-80.

[11] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis ofapproximations for maximizing submodular setfunctions", MathematicalProgramming, vol. 14, no. 1, 1978, pp. 265–294.

[12] J. H. Choi, "Living in Cyworld: Contextualising CyTies in South Korea,Use of Blogs Digital Formations", New York: Peter Lang, 2006, pp. 173-186.

[13] I. H. Lee, “On the convergence of informational cascades”, Journal of Economic theory, vol. 61, no. 2, 1993, pp. 395–411.

[14] J. R.Lee and C. W. Chung, “A Query Approach for Influence Maximization on Specific Users in Social Networks”, IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 2, Feb.2015, pp. 340-353.

[15] J. Tang, J. Sun, C. Wang, Z. Yang, “Social influence analysis in large-scale networks”, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, Paris, France, 2009, pp. 807–816.

[16] L .F. Feick and L. L. Price, “The Market Maven: A Diffuser of Marketplace Information”,Journal of Marketing, vol. 51,no. 1, 1987, pp. 83.

[17] L. L. Thurstone, “A Law of ComparativeJ udgment”, Psychological Review, vol. 34, 1927, pp. 273–86.

[18] M. A.Belch, K.A. Krentler, and L.A. Willis-Flurry, “Teen internet mavens: influence in family decision making”, Journal of Business Research, vol. 58, no. 5, 2005, pp. 569–575.

[19] M. Gladwell, “The tipping point: How little things can make a big difference”, Little, Brown, 2006.

[20] M. Forestier, A. Stavrianou, J. Velcin, and D. A. Zighed, “Roles in social networks: Methodologies and research issues”, Web Intelligence and Agent Systems: An international Journal, vol. 10, no. 1, 2012, pp. 117-133.

[21] M. Richardson and P. Domingos, “Mining knowledge-sharing sites for viral marketing”, in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002, pp. 61–70.

[22] N. Barbieri, F. Bonchi, and G. Manco, “Topic-Aware Social Influence Propagation Models”, Knowledge and information systems, vol. 37, no. 3, 2013, pp. 555–584.

[23] N. Ellison, C. Steinfield, and C. Lampe, “The benefits of Facebook friends: Social capital and college students use of online social network sites”, Journal of Computer Mediated Communication, vol. 12, no. 4, 2007, pp. 1143–1168.

[24] N. Garg and J. Koenemann, “Faster and simpler algorithms for multicommodity flow and other fractional packing problems”, SIAM Journal on Computing, vol. 37, no. 2, 2007, pp. 630–652.

[25] P. Domingos, M. Richardson, “Mining the Network Value of Customers”, Seventh International Conference on Knowledge Discovery and Data Mining, 2001.

[26] R. Abratt, D. Nel, and C. Nezer, "Role of the market maven in retailing: A general marketplace influencer", Journal of Business and Psychology, no. 1, 1995, pp. 31–55.

[27] R. Dorfman, P.A. Samuelson,and R.M. Solow , "Linear programming and economic analysis", Courier Corporation, 1958.

[28] R. Woodworth, "Experimental Psychology", New York:Henry Holt and Company, 1938.

[29] S. J. Barnes and A. D. Pressey, "In Search of the Meta-Maven", Psychology and Marketing, vol. 29, no. 3, 2012, pp. 167–185.

[30] T. K. Moon, "The expectation-maximization algorithm", Signal processing magazine, IEEE, vol. 13, no. 6, 1996, pp. 47–60.

[31] V. Brancaleone, J. Gountas, and others, "Personality characteristics of market mavens", Advances in Consumer Research, vol. 34, 2007, pp. 522.

[32] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks", in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 199–208.

[33] Y. Li,W. Chen, Y. Wang,and Z.L. Zhang, "Influence diffusion dynamics and influence maximization in : social networks with friend and foe relationships", in: Proceedings of the sixth ACM international conference on Web search and data mining (ACM), February 2013, pp. 657–666.

[34] "Graph API - Documentation - Facebook for Developers", Facebook Developers, 2016. [Online]. Available: https://developers.facebook.com/docs/graph-api. [Accessed: 11- Nov- 2016].