Chapter 1

Introduction

In the recent decade, the online social networks like Facebook, Flixster, Twitter, Google+, and Myspace, etc., provide a platform for user's interaction and communication, marketing and promotion for the new product, idea, and innovation. For example, people may tweet their opinions on breaking news; or may just update messages to tell friends what have happened in their daily life. Companies may hire influential users to promote new products such as movies and electronic goods. Besides, those information are flowing and can diffuse among users. Once users see something interesting, they can repost or forward these contents to their friends. If their friends also like the contents, they can further share them with their own friends, which thus causes information diffusion in the network, i.e., the so called effect of word-of-mouth [4, 5] spreading. Those users who adopt the information are called influenced or active.

However, how the information diffuses through network is usually unknown. Understanding the diffusion mechanism behind massive information is important for a wide range of applications like viral marketing [6], rumor control [7, 8], social recommendation [9], and revenue maximization [10], etc. This issue has attracted researchers from various fields including epidemiology, computer science, and sociology. There are different kind of diffusion models are proposed to describe and simulate this process, such as the independent cascade model (IC) [11], linear threshold model (LT) [11, 12], and epidemic models [12, 13]. Most models are contagious and assume that the information starts to diffuse from a source (seed) node set, and other nodes can access the information only from their neighbors.

The discovered diffusion models have been applied to many practical applications. For example, inspired by idea of viral marketing, the authors of [6] were first to study the influence maximization (IM) problem and introduced a probabilistic method to identifying influential users in the social network. In viral marketing, the advertising company plans to initially target a small number of influential users k of the network by giving them free samples of the product (the product is expensive or the company has limited budget so that they can only choose a small number of people). The company hopes that the initially selected users will recommend the product to their friends, their friends will influence their friend's friends and so on, thus many users will ultimately adopt the new product through the powerful word-of-mouth effect. Firstly, Kempe et al. [11] formatted the influence maximization problem and defined IM problem as to finds a set of seed users S of size k which maximize the influence spread in the social network, i.e.

$$S = argmax_{S^* \subseteq V \wedge |S^*|=k}\{\sigma(S^*)\}$$

where $\sigma(S)$ is an objective function to find the expected number of active nodes in the network after completion of the diffusion process. The integer k is a budget constraint which is very limited. The authors proved that influence maximization is NP-hard under traditional diffusion models.

1.1 Challenges and Issues

Influence maximization has significant importance in sociology, biology, physics, and computer science disciplines where systems are often represented as networks. Numerous techniques have been developed for influence maximization, yet the problem is not solved satisfactorily (see [3] for the reviews). Despite its immense application potential, influence maximization embraces enormous research challenges and issues, which are are discussed below.

1.1.1 Major Challenges

There are several challenges emerged along with the influence maximization problem, some of which are as follows:

- How to model the information diffusion process in a social network, which would heavily affect the influence spread of any seed set in influence maximization problem [11, 14].
- The influence maximization problem is theoretically complex in general. It has been proven that obtaining an optimal solution of influence maximization is NP-hard under most of the diffusion models [3, 11, 15]. Furthermore, due to the stochastic nature of information diffusion, even the evaluation of influence spread of any individual seed set is computationally complex. These theoretical results have shown that it is very challenging to retrieve a (near) optimal seed set and to scale to massive social graphs at the same time.
- Mostly information diffusion models as well as influence maximization algorithms are random in nature (see [3, 14] for the reviews). Different set of users are activated or influenced in different executions of algorithms for the same network. Accumulation of outputs obtained in different executions of an algorithm during performance evaluation is another challenging task.

1.1.2 Major Issues

There are several issues emerged along with the influence maximization problem, some of which are as follows:

- Identification of effective seed is a major issue in influence maximization problem. To alleviate this problem, several context-aware influence maximization techniques were introduced using some contextual features such as location [16, 17], time [18, 19], topic [20, 21], and competitive [22, 23], etc. These context-aware approaches generate effective seed with less efficiency and scalability.

- Most influence maximization algorithms are time inefficient and not scalable with large networks. Therefore, algorithms need to reduce the number of iteration, improve iteration complexity, and reduce search space, etc., to identify seed set efficiently. To alleviate this problem, several algorithms have been proposed such as heuristic metrics based [24–26], sub-modularity based [27–29], influence path based [30–32], community based [33, 34], etc. However, quality is still an issue.

- Assurance of both effectiveness and efficiency is a major issue during the selection of seed nodes (see [2] for the reviews). Measuring effectiveness incorporates topological and contextual information, whereas measuring efficiency involves in reducing search space, number of iterations, and complexity of iteration. The fundamental difference between the two measures has led to the trade-off between effectiveness and efficiency. Trade-off between effectiveness and efficiency is a major issue during performance evaluation of influence maximization algorithms.

- In real-world scenario, a company may intend to promote several competitive and non-competitive products in the same social network simultaneously [35–37]. In reality, in majority of social networks, different users have different interests for different products and have different acceptance probabilities of promotions from their social network friends. Therefore, compared to traditional influence maximization problem, the key issue here is how to decide the number of each product among the m items and identify the most k influential individuals to form the seed users.

- In real-world, there have been quite a number of users who maintain several accounts simultaneously, which allow them to propagate information across different networks [38–40]. Also, real world systems are complex as those cover wider range of aspects such as multiple relationships, organizational hierarchy, directional associations, etc., [41]. Therefore, identification of seed nodes in those diverse and multiple featured networks with single algorithm is a challenging task.

1.2 Objectives

The thesis is focused on five objectives that are discussed below in four categories.

1) Incorporating complex and non-monotone submodular diffusion models: Borodin et al. [42] states that certain diffusion models particularly those for investigating competitive influence in social networks may not be monotonic or submodular, and hence the original greedy approach cannot be used. This goal is achieved by either breaking the boundary of submodularity or developing and adopting more general techniques like heuristics and meta-heuristics. The submodularity of the influence function plays a vital role for designing efficient and theoretical bounded IM solutions. The submodularity of requirement of the influence function is too strict in certain scenarios. For example, the addition of some contextual feature like opinion [43] adopts non-submodular influence functions. The non-submodularity occurs as any node can switch between positive and negative opinions which are spread across the influence graph. Under such circumstances, the greedy framework is no longer effective. To provide a better solution than some simple heuristics, a possible direction is to model the influence function with more general functions, e.g., weakly submodular functions. Therefore, following objective is considered to adopt a broad family of diffusion models.

Objective 1: Develop an efficient influence maximization algorithm to adopt complex and broad family of diffusion models.

On the other hand, investigate the role of meta-heuristics like bio-inspired optimization techniques in influence maximization problem. Most IM algorithms yield at least quadratic complexity in the network where single connection exist among nodes. Often, meta-heuristic approaches such as bio-inspired evolutionary techniques are used for fast convergence. Hence, the motive is to design suitable objective function and to incorporate evolutionary technique for seed selection. Therefore, following objective is considered.

Objective 2: Investigate the role of evolutionary techniques in influence maximization.

2) Exploring the contextual features of influence maximization: In real-world scenario, users have their own interests (which can be represented as topics) and are more likely to be influenced by their friends (or friends of friends) with similar interests. For example, a cricket fan will more likely be influenced by famous cricket players rather than football players. The influence spread achieved varies depending on the type of information being propagated and the type of people in the network. The spread of influence can be improved by considering topical features in seed selection. Therefore, involving user's interest following objective is considered.

Objective 3: Develop an efficient and effective algorithm to explore topical feature in influence maximization.

3) Dealing with multiple products and multiple networks: In reality, in majority of social networks, different users have different interests for different products and thus have different acceptance probabilities of promotions from their social friends. Hence, the company may adopt a promotion strategy that each seed user can freely recommend several kind of items together and each non-seed user can consider accepting different categories promotions at the same time. On other hand, various kind of networks are developed to represent real system. On of the complex form of networks is multiple featured networks that consists several networks or relationships. Therefore, IM in such networks is challenging as multiple networks have to be processed. Most influence maximization algorithms yield at least quadratic complexity in the network where single connection exist among nodes. Hence, the motive is to design an efficient heuristic method to deal with multiple products for seed selection in multiple featured networks. Therefore, following objective is considered.

Objective 4: Develop an efficient influence maximization algorithm to handle multiple products and complex relationships in multiple networks.

4) Application of information diffusion and influence maximization: Once seed nodes are identified based on information diffusion in the networks, an immediate question may arise that how to utilize this information further in different applications. Obviously, information diffusion and influence maximization have different meaning from the viewpoint of the applications. Information diffusion provides an opportunity to users for propagating and receiving information to and from a region of influence which is beyond the scope of their social circles. Furthermore, this diffusion mechanism affects the formation of social links in the network. The communities can be identified based on label propagation and the nodes may influence the application based on their locality within the community they belong. Therefore, following objective is considered to identify missing links in networks based on information diffusion and influence maximization.

Objective 5: Examine applicability of information diffusion in the perspective of link prediction

1.3 Contributions

Main contributions of the thesis are divided into four parts addressing the aforementioned five objectives. Considering the first and second objectives an evolutionary algorithm is adopted which is not aligned with any specific diffusion models. A community-based context-aware influence maximization algorithm is proposed covering the third objective. Covering the multiple products and networks related objective i.e., fourth objective, a heuristic algorithm is proposed. Considering the fifth objective, applicability of information diffusion into link prediction problem is investigated and proposed an algorithm involving label propagation. Detail about the contributions are explained below.

1.3.1 Bio-inspired Optimization based Influence Maximization

To address the first and second objective, a learning-based influence maximization approach (LAPSO-IM) is proposed. Kempe et al. [11] stated that influence maximization problem is NP-hard under classical diffusion models. Discrete particle swarm optimization (DPSO) [44] is one of the most popular bio-inspired meta-heuristic for optimizing NP-hard problems due to its simplicity and efficiency. The major weakness of PSO is its premature convergence which leads to trapping in local optima. To maintain a trade-off between local and global search process, Hasanzadeh et al. [45] proposed a learning automata (LA) based DPSO named as DPSOLA. This work adopts and utilizes DPSOLA to study influence maximization problem. An objective function is defined by extending local influence evaluation function expected diffusion value (EDV) [46] to approximate within the two-hope area. A learning automata based PSO approach is presented to optimize local influence evaluation function by redefining the velocity and position rules based on learning automata. Empirical analysis on six real-world networks show that the proposed algorithm LAPSO-IM performs better than the base method DPSO [47] in terms of influence spread with almost same running time. LAPSO-IM is more time-efficient than base method IMLA [48] with the same level of influence spread. The statistical tests are also performed to demonstrate the significant difference between the proposed method LAPSO-IM and the state-of-the-art methods.

1.3.2 Context-aware Influence Maximization

Following the third objective, contextual feature of node's is considered in addition to community-based framework. Evidently, in real-world social networks, users have their own interests (which can be represented as topics) and are more likely to be influenced by their friends with similar interests. To address this problem, community based context-aware influence maximization (C2IM) algorithm is proposed. This work considers community structure and user's interests to find effective seed efficiently. First, the classical diffusion models are extended to utilize the contextual feature. The community-based framework is presented to select seed nodes efficiently. C2IM utilize node degree distribution to identify the community structure of the social network. To address user's interest in influence maximization problem, it models influence graph as a context-aware graph. Finally, empirical analysis on six real-world networks show that the proposed algorithm performs better than the base method CIM [49] in terms of influence spread with almost the same running time and more time-efficient than base method TIM [21] with approximate influence spread.

1.3.3 Multiple Influence Maximization in Multiple Featured Networks

Considering fourth objective, multiple products and multiple featured networks are considered for seed selection. Nearly most of the previous research works only focused on network topology and ignored some critical factors like multiple product advertisement, users engagement across networks, different channels of interaction, and network of networks etc. In real-world, distinct users have different influence or interest for distinct product in social networks. Therefore, each user has different influence probabilities from their neighbors. Nowadays, many users in networks are actively involved in multiple networks simultaneously. Therefore, overlapping users [1] play a pivotal role in information spreading across networks. Hence, the study of influence maximization problem in each network separately underestimates the social influence of overlapping users in other networks. This work proposes a novel framework multiple influence maximization across multiple social networks (MIM2) by considering the above scenario. The study proves that MIM2 problem is NP-hard and expected influence spread $\sigma(S)$ is sub-modular under traditional diffusion models. The experimental analysis are performed to validate the efficiency and influence spread of the proposed algorithm. It also covers the advantage of MIM2 framework over classical influence maximization problem.

[1] Users who actively engage in multiple networks simultaneously.

1.3.4 Applicability of Information Diffusion to Predict Missing Links

Prediction of missing links or future links in the network has great interest in several domains [50–52]. Considering the fifth objective, this work studies the applicability of information diffusion. Information diffusion provides an opportunity to users for propagating and receiving information to and from a region of influence which is beyond the scope of their social circles. Furthermore, this mechanism affects the formation of social links in the network. This work is primarily focused on two factors: information diffusion and importance of node's community. First, the proposed algorithm CLP-ID detects the community structure of the network using label propagation. Then, the algorithm utilizes the information about communities to estimate the likelihood score of missing links based on probabilistic model. An existing link is assigned with positive weight if associated nodes belongs to same community, otherwise assigned negative weight. Finally, empirical analysis on real-world networks validate the performance of proposed method against the state-of-the-art methods regarding different performance metrics.

1.4 Thesis Organization

The thesis is organized into seven chapters.

Chapter 2 provides literature survey on information diffusion models and influence maximization approaches.

Chapter 3 deals with submodularity of objective function. A learning based particle swarm optimization technique DPSOLA is adopted, and then used to identify seed nodes.

Chapter 4 deals with topical feature user's interest in influence maximization. A community-based framework is utilized to present a novel, efficient, and effective algorithm to find seed set. C2IM algorithm is designed involving the role of nodes in information diffusion and product adoption.

Chapter 5 deepens further the role of nodes in multiple product promotion and multiple featured networks. A novel MIM2 algorithm is proposed to deal with multiple influence maximization on multiple featured networks.

Chapter 6 studies prediction of missing links in the network as an application of information diffusion. A link prediction scheme using information diffusion perspective is proposed by incorporating the community structure.

Chapter 7 concludes the contributions and details possible future directions in respect of each of the proposed works.