Chapter 2
Diffusion of Information

In this chapter, we outline the techniques used in optimizing or facilitating information diffusion in social networks. We identify two problem definitions through which a broad survey of techniques in recent research is provided. Namely, we explore the problems of maximizing the spread of influence and minimizing the spread of misinformation in social networks. As different as these problems are in terms of the motivation behind them, they both rely on sub-problems that are very similar. Through our study of these two problems, we delve into more detail about the sub-problems: Sect. 2.2 model formation, Sect. 2.3 problem optimization, Sect. 2.4 large-scale data analysis, and Sect. 2.5 research trends.

2.1 Introduction

Diffusion of influence refers to circumstances where a point of view or behavior is widely spread in specific structures of propagation channels [35]. A diffusion can be associated with topological properties, such as scale, range, and temporal properties. This concept has been widely researched in the field of epidemiology, sociology, and marketing.

In early time, biology and epidemiology have conducted in-depth study on diffusion of virus within the group [8], and two classical models: SIS and SIR are proposed. In sociology and marketing area, research on diffusion focuses on the problems of innovation diffusion. In the early twentieth century, Schumpeter et al. [168] created innovative theory. Then the BASS model [3] opened up new research directions for this research area and derived a series of related models. Westerman et al. [202] studied the effect of system generated reports of connectedness on credibility and showed that there are curvilinear effects for the number of followers exist, such that having too many or too few connections results in lower judgments of expertise and trustworthiness. Lopez-Pintado et al. [120] studied the product

W. Xu, W. Wu, *Optimal Social Influence*, SpringerBriefs in Optimization,
https://doi.org/10.1007/978-3-030-37775-5_2

diffusion in complex social networks. He considered the mutual influence among individuals on the micro-level into the propagation equation based on mean-field theory and found out that innovation diffusion in complex networks has a threshold which is closely related to the degree distribution and propagation functions of the network.

Understanding, capturing, and being able to predict influence diffusion can be helpful for several areas such as viral marketing, cyber security, and Web search. For instance, if we consider the case of marketing, it may be useful to know which are the features that control the process of diffusing information when it is created to, e.g., better advertise a product or to better protect it against attacks on the network. The marketing may also benefit from information such as how many initial users to start with in a marketing campaign (budget optimization), how much time to wait between actions, etc. In the case of security, criminal investigators generally need to understand the information flow between, e.g., members of a given community to extract hints regarding possible guilt or innocence of a person or a group of persons. This is clearly an observation phase where the user wants to understand the route that information took and possible links. Finally, as Web search evolves, if we consider the case of subscriptions to feeds, a propagation prediction model may be useful for the user to, e.g., subscribe to the most interesting topic according to its expected growth (in addition to his interests). This reflects a more active usage of the diffusion prediction.

2.2 Model Formation

Central to optimization problems relating to information diffusion is the problem of identifying the right diffusion model. Therefore, we provide a survey of available models and address the following questions: What are the necessary and sufficient parameters of an accurate model? How can we validate the use of a specific model? How can one obtain data about the parameters? Given the intricacy of human interactions, finding the right diffusion model is still an open problem, even in the presence of the large datasets available today. In this section, we give an overview of the most common propagation models, including epidemic models [78, 85], the "Bass" model [16] for product adoption, and basic diffusion models such as independent cascade (IC) and linear threshold (LT) [56]. The goal is also to learn about fundamental properties of such processes in a variety of settings.

2.2.1 Epidemic Models

Infectious agents have had decisive influences on the history of mankind. Fourteenth century Black Death has taken lives of about a third of Europe's population at the time. The first major epidemic in the USA was yellow fever epidemic in

Philadelphia in 1793, in which 5000 people died out of a population of 50,000. This epidemic has had a major impact on the life and politics of the country. Thucydides describes the Plague of Athens (430–428 BC): 1050 of 4000 soldiers on an expedition died of a disease. Thucydides gives a detailed account of symptoms: some so horrendous that the last one—amnesia—seems a blessing. An interesting feature of this account is that there is no mention of person-to-person contagion, which we now suspect with most new diseases. It was not until the nineteenth century that the person-to-person contagion on was beginning to be discussed. In this book, we will mostly be interested in modeling infectious diseases, where the major means of disease spread comes from the person-to-person interaction.

The practical use of epidemic models must rely heavily on the realism put into the models. This does not mean that a reasonable model can include all possible effects, but rather incorporate the mechanisms in the simplest possible fashion so as to maintain major components that influence disease propagation. Great care should be taken before epidemic models are used for prediction of real phenomena. However, even simple models should, and often do, pose important questions about the underlying mechanisms of infection spread and possible means of control of the disease or epidemic.

We begin with classical papers by Kermack and McKendrick (1927, 1932, and 1933). These papers have had a major influence on the development of mathematical models for disease spread and are still relevant in many epidemic situations. The first of these papers laid out a foundation for modeling infections which, after recovery, confer complete immunity (or in case of lethal diseases—death). The population is taken to be constant—no births or deaths other than from the disease are possible—consistent with the course of an epidemic being short compared with the life time of an individual. If a group of infected individuals is introduced into a large population, a basic problem is to describe the spread of the infection within the population as a function of time. In the course of time the epidemic may come to an end. One of the most important questions in epidemiology is to ascertain whether this occurs only when all of the initially susceptible individuals have contracted the disease or if some interplay of infectivity, recovery, and mortality factors may result in epidemic "die out" with many susceptibles still present in the unaffected population.

Mathematical modeling of infectious diseases is a tool to investigate the mechanisms for outbreak and spread of diseases and to predict the future course in order to control an epidemic. Generally there are several types of epidemic models.

First, stochastic models. The epidemic process has random nature. Stochastic models are used to estimate the probabilistic quantities for the outcome events, such as the probability distribution of extinction time, the probability distribution of final epidemic size, the associate mean, and so on.

Second, deterministic compartmental models. The transition rate from one class (compartment) to the other one is characterized by derivative mathematically. If we assume that the population size is differentiable with respect to time, in the limiting of large population, the time evolution of behavior of each subgroup can be approximated by the deterministic dynamics.

In the category of deterministic compartmental models, there are two classical classes: SIR and SIS. In SIS and SIR epidemic models, individuals in the population are classified according to disease status, either susceptible, infectious, or immune. Healthy ("S" = susceptible) nodes become sick ("I" = infected) stochastically from their infected neighbors with a probability. Alternatively, a sick node becomes healthy ("R" = removed) and open to re-infection with a probability. These two parameters are also referred to as the birth rate and death rate of the virus.

The tipping point, or epidemic threshold, of an SIS epidemic model is the condition under which an infection will die out exponentially quickly irrespective of initial infection, as opposed to spreading out, causing and epidemic. For a survey on SIS and numerous other epidemic models, please refer to Hethcote [85].

2.2.2 Product Adoption Model

The well-known first purchase diffusion models in marketing are those of Bass [16], Fourt and Woodlock [67], and Mansfield [129]. These early models attempted to describe the penetration and saturation aspects of the product diffusion process.

The main impetus underlying diffusion research in marketing is the Bass model. Subsuming the models proposed by Fourt and Woodlock [67] and Mansfield [129], the Bass model assumes that potential adopters of an innovation are influenced by two means of communication—mass media and word of mouth. In its development, it further assumes that the adopters of an innovation comprise two groups. One group is influenced only by the mass media communication (external influence) and the other group is influenced only by the word-of-mouth communication (internal influence). Bass termed the first group "Innovators" and the second group "Imitators." Unlike the Bass model, the model proposed by Fourt and Woodlock [67] assumes that the diffusion process is driven primarily by the mass media communication or the external influence. Similarly, the model proposed by Mansfield [129] assumes this process is driven by word of mouth.

Figures 2.1 and 2.2 are plots of conceptual and analytical structure underlying the Bass model. As noted in Fig. 2.1, the Bass model conceptually assumes that "Innovators" or buyers who adopt exclusively because of the mass media communication or the external influence are present at any stage of the diffusion process. Figure 2.2 shows the analytical structure underlying the Bass model. As depicted, the noncumulative adopter distribution peaks at time $T*$, which is the point of inflection of the S-shaped cumulative adoption curve. Furthermore, the adopter distribution assumes that an initial pm (a constant) level of adopters buy the product at the beginning of the diffusion process. Once initiated, the adoption process is symmetric with respect to time around the peak time T^* up to $2T^*$. That is, the shape of the adoption curve from time T^* to $2T^*$ is the mirror image of the shape of the adoption curve from the beginning of the diffusion fusion process up to time T^*. In general, the Bass model is a popular model appeared at an early stage for product adoption. For more information, please refer to [128].

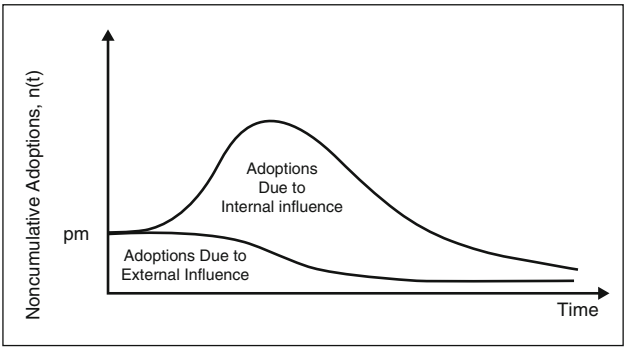

Fig. 2.1 Adoptions due to external and internal influences in the bass model

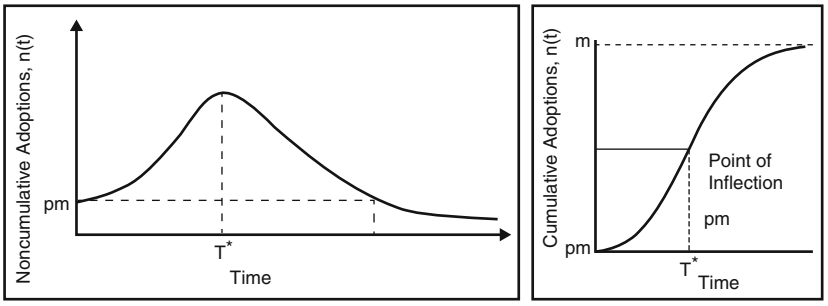

Fig. 2.2 Analytical structure of the bass model

2.2.3 Diffusion Models

Diffusion is the process by which information passes from neighbor to neighbor [137]. Real-world examples include viral marketing, innovation of technologies, and infection propagation. Diffusion models are the framework on which diffusion occurs.

Definition 2.1 A diffusion model is a graph $G = V, E$ along with a collection of activation functions $F = (f_v)_{v \in V}$, where f_v is a $\{\emptyset, \{v\}\}$ valued function on $2^{|V|}$.

The output of a function f_v is a random variable based on the activation function.

Vertices on this graph are usually individuals and the activation function models the influence individuals exert on others. The activation function usually depends only on the neighbors of v, denoted $N(v)$. This means that $f_v(S) = f_v(N(v) \bigcap S)$.

Definition 2.2 Diffusion is the process on a diffusion model M, $S = (S_t)_{t=0}^{n-1}$ started at $S \subseteq V$:

1. set $S_0 = S$
2. for $t > 1$ set $S_t = f(S_{t-1}) = def \bigcup_{v \in V} f_v(S_{t-1})$

The set of nodes activated at the end of diffusion is denoted as $\sigma(S) = \bigcup_{t=0}^{n}(S_t)$.

Diffusion occurs in time steps t. At each time step, all previously activated nodes remain activated and individuals are either activated or deactivated based on the activation functions. Diffusion can run on a fixed number of time steps or indefinitely. Diffusion is said to have stopped when the set of activated nodes in time step t_k is the same as the set in time step t_{k+n} for all $n \geq 1$.

One class of diffusion models, namely threshold model, adds an influence threshold to each individual, which, when overcome, triggers the individual to be activated. There is a cumulative effect of these models, as it takes a critical number of influential neighbors to activate an individual.

2.2.3.1 General Threshold Model

This model was defined by Kempe et al. [94] and Mossel and Roch [136].

Definition 2.3 The general threshold model is a diffusion model with

1. A set of threshold values $(\theta_v)_{v \in V}$, where θ_v is in the range [0, 1].
2. Node v being activated if $f_v(S) \geq \theta_v$, where S is the set of neighbors of v.

The activation function on the general threshold model depends on the activated neighbors of v. There is an assumption of monotonicity on this model made to reflect that adding active neighbors to a node increases likelihood of the node being activated.

Definition 2.4 A function $f : 2^V \rightarrow R$ is monotone if $f(S) \leq f(T)$ for all $S \subseteq T \subseteq V$.

This property captures that activating more nodes will always have an increasing effect on the nodes that will be activated at a future time.

2.2.3.2 Linear Threshold Model

The linear threshold model is a specialized form of general threshold models. The linear threshold model, LT model in short, is more often used in marketing research.

Definition 2.5 The linear threshold model is a diffusion model with all of the properties of the general threshold model with

1. A set of weights $(p(u, v))_{(u,v) \in E}$ with the property $\Sigma_{u \in N(v)} p(u, v) \leq 1$.
2. Activation function of the form $f_v(S) = \sum_{u \in N(v)} p(v, u)$ with $f(\emptyset) = 0$.

Cascade models of diffusion give each individual the ability to influence their neighbors as soon as they are activated. This is opposed to the threshold models that rely on a cumulative effect. This model has the property that the more nodes that have attempted to influence a node, the less likely the node is to be activated.

Here we give a definition of a specialized cascade model, namely the independent cascade model, IC model in short.

Definition 2.6 The **independent cascade model** is a diffusion model with the following properties:

1. Each arc (u, v) has associated the probability $p(u, v)$ of u influencing v.
2. Time unfolds in discrete steps.
3. At time t, nodes that became active at $t - 1$ try to active their inactive neighbors and succeed according to $p(u, v)$.

Note that the probability of a node u influencing a node v is independent of the set of nodes S that has attempted to influence v.

There is an assumption of monotonicity on this model made to reflect that adding active neighbors to a node increases likelihood of the node being activated.

2.2.3.3 History-Sensitive Cascade Model

The history-sensitive cascade model, designed by Foster and Potter, is essentially a reformat of the linear threshold model and is not a different diffusion model itself. In their research into the spread of influence, Foster and Potter propose the idea that the probability of a node being activated increases the longer the node is in contact with other activated nodes. Since at every time step more neighbors can be added, while the combined influence never goes down, the probability that any node is activated increases with each new neighbor added. This reflects the monotonic property of the linear threshold model.

Foster and Potter studied the exact effects of diffusion over time on the probability that any node would be activated at time step k. They studied this effect on tree-structure graphs and also on general graphs and proposed algorithms for determining these probabilities. To attain the probability of a node being activated at any given time step, a Markov chain model is used.

Definition 2.7 A Markov chain is a sequence of random variables X_1, X_2, X_3, with the property that $Pr(X_{n+1} = x | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$.

A Markov chain is a collection of states with transitions between states such that the probability of transitioning to any state from any other state depends only on the current state. Foster and Potter use a Markov chain model that encode sets of active nodes in binary strings and then create a transition matrix that maps the probability of transitioning from any set of activated nodes to any other set. By iterating over this transition matrix, it is possible to find the exact probability of any node being activated at any time step for any arbitrary graph.

2.2.3.4 Cascade Models

Cascade models of diffusion give each individual the ability to influence their neighbors as soon as they are activated. This is opposed to the threshold models that rely on a cumulative effect. This model has the property that the more nodes that have attempted to influence a node, the less likely the node is to be activated.

2.2.3.5 General Cascade Model

This model was designed by Kempe et al. [94] as a general form of the cascade model. This model has the property that the more nodes that have attempted to influence a node, the less likely the node is to be activated.

Definition 2.8 The general cascade model is a diffusion model with the following properties:

1. nodes are live at time t if they were activated in time $t-1$.
2. a collection of probability functions $P = p_v, v \in V$ where p_v is a [0, 1]-valued function on 2^V.
3. activation function of the form

$$f_v(W) = \begin{cases} 1 & \text{with probability } p_v(W) \\ 0 & \text{otherwise} \end{cases}$$

   where $W \subseteq S$ and every $w \in W$ is live at time t.
4. node v being activated in time t if $f_v(W) = 1$, where W is the set of neighbors of v live at time t.
5. the order-independence property, defined below.

Note that each of the following definitions use p_v as an element of P and are defined over all $v \in V$. Likewise for f_v as an element of F defined over all $v \in V$.

Definition 2.9 The order-independence property states that when $\sigma : 1, \ldots, r \rightarrow 1, \ldots, r$ is a permutation function and $u_1, \ldots, u_r$ and $u_{\sigma_1}, \ldots, u_{\sigma_r}$ are two permutations of T, and $T_i = u_1, \ldots, u_{i-1}$ and $T_i' = u_{\sigma_1}, \ldots, u_{\sigma_{i-1}}$, then

$$\prod_{i=1}^{r}(1 - p_v(u_i \bigcup S \bigcup T_i)) = \prod_{i=1}^{r}(1 - p_v(u_{\sigma_i} \bigcup S \bigcup T_i'))$$

for all sets S disjoint from T.

The probability of a node u influencing a node v depends on the set S of nodes that has already attempted to influence v. However, the ordering dependence property states that the probability of u activating v does not depend on the order of nodes in the set S that have previously attempted to activate v.

2.2.3.6 General Cascade and General Threshold Equivalence

The general cascade model has been shown to be equivalent to the general threshold model [94] under the following mapping:

1. for the probability function in the general cascade model:

$$p_u(u \bigcup S) = \frac{f_v(S \bigcup u) - f_v(S)}{1 - f_v(S)}$$

2. for the activation function in the general threshold model:

$$f_v(S) = 1 - \prod_{i=1}^{r}(1 - p_v(u_i) \bigcup S_{i-1})$$

where $S = u_1, \ldots, ur$ and $S_i = u_1, \ldots, u_i$.

This effectively says that by choosing the edge weights in either model, an instance of the general threshold model may be transformed into an instance of the general cascade model. This mapping ties the two models together and shows that diffusion is an equally hard problem on either model. Therefore, conclusions on one model also apply to the other model.

2.2.3.7 Decreasing Cascade Model

The decreasing cascade model was also defined by Kempe et al. [94] and is an extension of the general cascade model with the property that the more nodes that have attempted to activate a node, the less probability there is that the node becomes activated.

Definition 2.10 The decreasing cascade model is a diffusion model with all of the properties of the general cascade model with the additional property where $p_v(u \bigcup S) \geq p_v(u \bigcup T)$ whenever $S \subseteq T$.

2.2.3.8 Independent Cascade Model

This model was initially investigated by Goldenberg et al. in the context of marketing [73] and was defined by Kempe et al. [93]. Along with the linear threshold model, this model is classically used for studying diffusion on networks. It exists as a special case of the decreasing cascade model.

Definition 2.11 The independent cascade model is a diffusion model with all of the properties of the decreasing cascade model with the additional property that the $p_v(u \bigcup S) = p_v(u)$ for all sets $S \subseteq V$.

This means that the probability of a node u influencing a node v is independent of the set of nodes S that has attempted to influence v. Since we will be using this model for the remainder of our research, it is helpful to define some shorthand. We can look at this model as a set of edge probabilities on a graph.

Definition 2.12 On the independent cascade model, an edge probability, $b_{u,v}$ is the probability that a node u has to infect v whenever u is infected.

Note that $b_{u,v}$ does not necessarily equal $b_{v,u}$ and in fact, it will be the case in certain situations in our research that if $b_{u,v}$ is non-zero, that $b_{v,u}$ is 0.

It should be noted that the independent cascade model has the property that a node has exactly one time step in which it is infected to infect other nodes. That is, each node is infectious for exactly one time step and then can no longer be infected, nor can it infect any other nodes.

2.3 Problem Optimization

To better understand the underlying ideas behind diffusion and social networks, we study the formulations and optimizations for two important problems in social networks: (1) maximizing the spread of influence and (2) limiting the spread of misinformation, which is also called rumor blocking in some related work.

To begin with, we will cover some basic knowledge of social network. Social network is modeled as a directed graph $G = (V; E)$ with vertices in V modeling the individuals and edges in E modeling the relationship between individuals. For example, in co-authorship graphs, vertices are authors of academic papers and two vertices have an edge if the two corresponding authors have coauthored a paper.

2.3.1 Influence Maximization

An intensively studied problem in viral marketing is that, by picking a small group of influential individuals in a social network—say, convincing them to adopt a product—it will trigger the largest cascade of influence by which many users will try the product ultimately. Domingos and Richardson [53] are the first to pose it as a algorithmic problem and solve it as a probabilistic model of interaction. In [93], Kempe et al. formalize it as the problem of influence maximization.

A social network is modeled as a directed graph $G = (V, E)$ with vertices in V modeling the individuals and edges in E modeling the relationship between individuals. For example, in co-authorship graphs, vertices are authors of academic papers and two vertices have an edge if the two corresponding authors have coauthored a paper. Let p denote the influence probabilities between two vertices. The influence is propagated in the network according to a diffusion model m. Let S be the subset of vertices selected to initiate the influence propagation, which is also

called seed set. Let $\sigma_m(S)$ be the expected number of influenced nodes at the end of propagation process. The formal definition of influence maximization problem is given as follows:

Problem 2.1 (Influence Maximization) Given a directed and edge-weighted social graph $G = (V, E, p)$, a propagation model m, and an integer $k \leq |V|$, find a seed set $S \subset V, |S| = k$, such that the expected influence $\sigma_m(S)$ is maximum.

This problem is also referred to as the identification of influential users or opinion leaders in a social network. This problem under both independent cascade (IC) and linear threshold (LT) propagation models is shown to be NP-hard [94], and so attempts have been made at approximating the value of $\sigma_m(S)$.

For a diffusion model with a non-negative, monotone submodular activation function, a greedy hill-climbing algorithm approximates the optimum within a factor of $(1 - 1/e) - \epsilon$ for any real number ϵ, as shown by Kempe et al. [93]. By greedy hill-climbing algorithm we mean an algorithm which, at every step, adds to the output set the node that currently has the highest influence spread. The challenge of the greedy algorithm rises when selecting a new vertex ν that provides the largest marginal gain $\sigma_m(S + \nu) - \sigma_m(S)$ compared to the influence spread of current seed set S. Computing the expected spread given a seed set turns out to be a difficult task under both the IC model and the LT model. Instead of finding an exact algorithm, Kempe et al. run Monte Carlo simulations of the propagation model for sufficiently many times (10,000 trials) to obtain an accurate estimate of the influence spread, leading to a very long computation time.

A vast number of papers have studied improving the efficiency and availability of the influence maximization [25, 37, 39, 130, 149, 183, 187, 188]. In [37], Chen et al. also propose a degree discount heuristics with influence spreads and combines a Cost-Effective Lazy Forward (CELF) scheme to further improve the greedy algorithm. In [39], Chen et al. propose a scalable heuristic called DAGs (local directed acyclic graphs) for the linear threshold model. They construct local DAGs for each node and computing the expected spread over DAGs can be done in linear time while over general graphs it is #P-hard. In [130], Mathioudakis et al. simplified the network to accelerate the speed of finding seeds. However, these heuristics lack of theoretical guarantees. At this front, the state of the art is the reverse influence sampling (RIS) approach [25, 188]. These methods attempt to generate a $a1 - 1/ - \epsilon$ approximation solution with minimal number of RIS samples. And the IMM algorithm [188] is among the most competitive ones so far. In [149], Nguyen et al. generalize the RIS sampling methods into sampling frameworks and optimize it by an innovative stop and share strategy. Their method uses minimum number of samples while meeting strict theoretical thresholds for the influence maximization problem.

Another issue for Kempe's method is that it assumes a weighted social graph as input and does not address the problem of learning influence probabilities. In [164], Saito et al. study how to learn the probabilities of the IC model from a set of past propagations by formalizing this as a likelihood maximization problem and then applying the expectation maximization (EM) algorithm to solve it; Goyal et al. [75,

76] propose a credit model for learning influence probability from pure historical action logs which takes the temporal nature of influence into account. In [207], Xu et al. first present a method to identify influential entities in large social networks based on a weighted maximum cut framework which is totally separate from traditional method of greedy strategy while maintaining high efficiency. Moreover, they have developed a new method of learning influence strength by analyzing both social network structure and historical user data.

Some variations are proposed to handle different real-world requirements, such as looking at communities, competitive and complementary influence maximization. Leskovec et al. [115, 183] optimized placements for a set of social sensors such that the propagation of information or virus can be effectively detected in a social network. Lappas et al. [111] discover a set of key mediators which determine the bottlenecks of influence propagation if seed nodes try to activate some target nodes. Sun et al. [183] study the multi-round influence maximization problem, where influence propagates in multiple rounds independently from possibly different seed sets.

A characteristic common to the studies discussed so far is the assumption that information cascades of campaigns happen in isolation. Next we introduce a group of problem formulations that capture the notion of competing campaigns in a social network [19, 26, 33, 40, 104, 190]. This scenario will frequently arise in the real world: multiple companies with comparable products will vie for sales with competing word-of-mouth cascades; similarly, many innovations face active opposition also spreading by word of mouth. Carnes et al. [33] study the strategies of a company that wishes to invade an existing market and persuade people to buy their product. This turns the problem into a Stackelberg game where in the first player (leader) chooses a strategy in the first stage, which takes into account the likely reaction of the second players (followers). In the second stage, the followers choose their own strategies having observed the Stackelberg leader decision, i.e., they react to the leader's strategy. Carnes et al. use models similar to the ones proposed in [93] and show that the second player faces an NP-hard problem if aiming at selecting an optimal strategy. Furthermore, the authors prove that a greedy hill-climbing algorithms leads to a $(1 - 1/e - \epsilon)$-approximation.

Around the same time, Bharathi et al. [19] introduce roughly the same model for competing rumors and they also show that there exists an efficient approximation algorithm for the second player. Moreover they present an FPTAS for the single player problem on trees. Kostka et al. [104] considered the rumors diffusion as a game theoretical problem under a much more restricted model compared with IC and LT. They showed that the first player did not always obtain benefit although he/she started earlier. Trpevski et al. [190] propose a competitive rumors spreading model based on SIS model in epidemic domain, but they did not address the issue of influence maximization or rumor blocking. Borodin et al. in [26] study competitive influence diffusion in several different models extended from LT. Chen et al. [40] address positive influence maximization under an extension of the IC model with negative opinions about the product or service quality.

2.3.2 Misinformation Minimization

While the ease of information propagation in social networks can be very beneficial, it can also have disruptive effects. A number of examples of this sort are the spread of misinformation on swine flu in Twitter [135], exaggerated reports on a bomb attack in Grand Central, and celebrities that are falsely claimed as being dead [86]. We specifically focus on the study that addresses the problem of influence limitation [30] where a bad campaign starts propagating from a certain node in the network and use the notion of limiting campaigns to counteract the effect of misinformation. The problem of misinformation minimization can also be called as rumor blocking problem or influence limitation problem. Its definition is defined as follows.

Problem 2.2 (Misinformation Minimization) Given a graph $G = (V; E; p)$, where p represents its positive and negative edge weights, a negative seed set N_0, and a positive integer k, the goal is to find a positive seed set S of size at most k such that the expected number of negatively activated nodes is minimized, or equivalently, the reduction in the number of negatively activated nodes is maximized.

Kimura et al. in [98] deal with influence limitation problem through blocking a certain number of links in a network. The most recent works regarded with this problem include [30, 84, 147]. In [30], Budak et al. study the controlling of negative information in social networks, that is, when negative information is diffused in networks, how to select some nodes to implant positive information in order to correct the information attitude in the whole network to a maximizing extent. They prove that under an extension of the IC model, the eventual influence limitation (EIL) problem is NP-hard. They also examine a more realistic problem of influence limitation in the presence of missing information and introduced an algorithm called predictive hill-climbing approach which has good performance.

In [84], He et al. propose a competitive linear threshold (CLT) model to address the influence blocking maximization (IBM) problem, which is an extension to the classic linear threshold model. They prove that this problem under CLT model was submodular and theoretically obtained a $(1 - 1/e)$-approximation ratio by a greedy strategy. To improve the efficiency, they further propose the CLDAG algorithm that is similar to the LDAG algorithm in [39]. In [147], a β_T^I-Node Protector problem is proposed by Nguyen et al., which is actually the extensions of the misinformation minimization problem under LT and IC models. The goal is to find the smallest set of highly influential nodes that can limit the viral spread of misinformation originated from set I to a desired rate $(1 - \beta)$ $(\beta \in [0, 1])$ in T time steps. They present a greedy viral stopper (GVS) algorithm that greedily adds nodes with the best influence gain for β Node Protectors to the current solution. They also apply GVS to the network restricted to T-hop neighbors of the initial set I and reached a slightly better bound for β_T^I-Node Protector problems. Besides, a community based algorithm which returns a good selection of nodes to decontaminate in a timely manner is proposed.

2.4 Large-Scale Data Analysis

No matter which technique is used in studying information diffusion, large-scale data analysis is a significant aspect of study as well as being a significant challenge. In this part, we will introduce several representative data analysis techniques used in the social influence analysis. With the increase of studies in social networks, there are a number of datasets available to researchers [109, 113, 146].

As data grows, data mining and machine learning applications start to embrace the Map-Reduce paradigm, e.g., news personalization with Map-Reduce EM algorithm [49], Map-Reduce of several machine learning algorithms on multicore architecture [45]. For the networking data, graphical probabilistic models are often employed to describe the dependencies between observation data. Markov random field [180], factor graph [105], restricted Boltzmann machine (RBM) [201], and many others are widely used graphical models. In [186], Tang et al. proposed a topical factor graph (TFG) model, for quantitatively analyzing the topic based social influences. Compared with the existing work, the TFG can incorporate the correlation between topics. They also proposed a very efficient algorithm for learning the TFG model. In particular, a distributed learning algorithm has been implemented under the Map-Reduce programming model.

The techniques used in Web community discovery can also be applied in social influence analysis. The problem of detecting such communities within networks has been well studied. Early approaches such as spectral partitioning, the Kernighan-Lin algorithm, hierarchical clustering, and G-N algorithm work well for specific types of problems (particularly graph bisection), but perform poorly in real networks. Recently, most works focus on graph partitioning approaches. The most popular partition technique in the literature is k-means clustering, which aims to separate the nodes in k clusters such to maximize/minimize a given cost function based on distances between nodes and/or from nodes to centroids. In [209], Q. Yan et al. proposed a two-phase method that combines community detection with naive greedy algorithm to improve time efficiency of influence maximizing problem with multiple spread model. In the first phase, they use efficient clustering algorithm such as kernel k-means to partition graph nodes into k clusters, with the parameter k related to the number of influential nodes. In the second phase, in each community, they apply techniques in social influence maximization to find influential nodes in each cluster. Similar work has [48].

2.5 Research Trends

Social networks provide large-scale information infrastructures for people to discuss and exchange ideas about different topics. The general problem of network influence analysis represents a new and interesting research direction in social network mining. There are many potential future directions of this work. Even though the

influence diffusion in social networks has been intensively studied, we note that there are three essential dimensions emerging from the analysis we performed, which could be of great benefits for future researchers.

2.5.1 Learn Influence Probabilities of Diffusion Models

In social network analysis, two information diffusion models: the independent cascade (IC) and the linear threshold (LT) are widely used to solve such problems as the influence maximization problem and the misinformation minimization problem. These two models focus on different information diffusion aspects. The IC model is sender-centered (push type) and each active node independently influences its inactive neighbors with given diffusion probabilities. The LT model is receiver-centered (pull type) and a node is influenced by its active neighbors if their total weight exceeds the threshold for the node. What is important to note is that both models have parameters that need be specified in advance: diffusion probabilities for the IC model, and weights for the LT model. However, their true values are not known in practice. This poses yet another problem of estimating them from a set of information diffusion results that are observed as time sequences of influenced (activated) nodes. This falls in a well-defined parameter estimation problem in machine learning framework.

In [165], K. Saito et al. extended both IC and LT models to be able to simulate asynchronous time delay. They learned the dependency of the diffusion probability and the time delay parameter on the node attributes by solving a formulated problem named as the maximum likelihood estimation problem, and an efficient parameter update algorithm that guarantees the convergence is derived. Other efforts of learning parameters of the influence graph from history data include the work [75, 162]. In [75], A. Goyal et al. proposed both static and time-dependent models for capturing influence. Moreover, they presented optimized algorithms for learning the parameters of the various models based on social networks and historical action logs.

2.5.2 Learn the Speed of Influence Spread in Networks

It has been observed that information spreads extremely fast in social networks. There has been some but not enough theoretical results about the analysis of influence spread speed. In [52], B. Doerr et al. have shown that for preferential attachment graphs the classic push-pull strategy needs $\Theta(logn)$ rounds to inform all vertices. The slightly improved version which avoids that a vertex contacts the same neighbor twice in a row only needs $\Theta(logn/loglogn)$ rounds, which is best possible since the diameter is of the same order of magnitude. In [66], N. Fountoulakis et al. establishes for a class of random graphs ultrafast time bounds on the running time of

the synchronous push-pull protocol that is needed until the majority of the vertices are informed. They present the first theoretical analysis of this protocol on random graphs that have a power-law degree distribution with an arbitrary exponent $\beta > 2$. Their main findings reveal a striking dichotomy in the performance of the protocol that depends on the exponent of the power law. More specifically, it is shown that if $2 < \beta < 3$, then the rumor spreads to almost all nodes in $\Theta(loglogn)$ rounds with high probability. On the other hand, if $\beta > 3$, then $\Theta(logn)$ rounds are necessary.

2.5.3 Study Variations of Influence Maximization

Traditional diffusion models including IC and LT do not fully incorporate important temporal aspects that have been well observed in the dynamics of influence propagation. Firstly, the propagation of influence from one person to another may incur a certain amount of time delay, which is obvious from recent studies by statistical physicists on empirical social networks. Secondly, the spread of influence may be time-critical in practice. In a certain viral marketing campaign, a company might wish to trigger a large cascade of product adoption in a fairly short time frame, e.g., a 3-day sale. Therefore it is very meaningful to extend the influence maximization problem to have a time constraint.

Chen et al. [41] proposed the time-critical influence maximization problem, in which one wants to maximize influence spread within a given deadline. In their model influence delays are constrained to follow the geometric distribution. In [119], B. Liu et al. proposed a new problem of the time constrained influence maximization in social networks based on a latency aware independent cascade model. They also proposed to use influence spreading paths to quickly and effectively approximate the time constrained influence spread for a given seed set. Sun et al. [183] propose to study multi-round influence maximization problem, which models the viral marketing scenarios in which advertisers conduct multiple rounds of viral marketing to promote one product.