Chapter 2

Literature Review

In this Chapter, we present an overview of IM approaches which contains an introduction to the research so far. Inspired by the idea of viral marketing, Pedro and Matt [6] were first to introduced IM as an optimization problem in 2001. Formally, IM problem is formatted by Kempe et al. [11] in 2003. Let us assume that a social network $G(V,E,W)$ where V and E denote individuals and their relationship (undirected/ directed) in the network. W is the measure of tie strength. The objective of IM problem is to select k most influential seed users. The influence propagation of any active node is dependent on diffusion model.

Before the formalization of IM problem definition, we present some definitions to better understanding of problem. Some basic definitions and IM problem formalization are given as follow.

Definition 2.0.1. *(Social Network).* A social network with N users and M social ties is represented as a weighted-directed graph $G(V,E,W)$. Here,

V denotes set of users, $|V| = N$, and E represents a set of relationships, $|E| = M$, and W represents edge-weight. A social network is also known as an influence graph.

Definition 2.0.2. *(Neighbors).* Neighbors $N(u)$ of node u is defined as the set of users v such that $v \in N(u)$ iff $\exists (u,v) \in E$, $v \in V$. $N_{inc}(u)$ and $N_{out}(u)$ denote in and out neighbors of node u respectively.

Definition 2.0.3. *(Degree centrality).* Degree centrality is defined as the number of links incident upon a node i.e. $C_D(u) = |N(u)|$. In directed social network, degree of u is considered as $C_D(u) = |N_{out}(u)|$.

Definition 2.0.4. *(Seed nodes).* Seed nodes (S) are the set of nodes who act as the source of the information propagation process in the social network, $|S| = k$, $S \in V$.

Definition 2.0.5. *(Active Node).* A node $u \in V$ is called active if either $u \in S$ or u adopted the information propagated by previously active nodes $v \in V_A$ under diffusion model. Once u is activated, then $V_A \leftarrow \{V_A \cup u\}$.

Definition 2.0.6. *(Influence spread).* Influence spread $I_S(S)$ of the seed set S is defined as the number of active users after diffusion process under a diffusion model, i.e., $I_S(S) = |V_A(S)|$.

Definition 2.0.7. *(Information Diffusion Model (IDM/DM)).* Given an influence graph $G = (V,E,W)$, a seed set $S \subseteq V$ and an information diffusion model captures the stochastic process for S influence spreading on graph G.

Definition 2.0.8. *(Influence Maximization (IM) [11]).* Given an influence graph $G = (V, E, W)$, an information diffusion model, a positive integer k, then influence maximization process selects a seed set $S \subseteq V$ of k users to maximize the influence spread in G, i.e., $\sigma(S) = argmax_{S^* \subseteq V \wedge |S^*|=k} \sigma(S^*)$.

The objective function $\sigma(S)$ is dependent on diffusion model and estimates the expected adoption of the product in the network. IM problem is also known as the k-node subset selection problem. Kempe et al. proved that the objective function $\sigma(S)$ is sub-modular under classical diffusion models. The authors pointed out that IM problem is *NP-hard*. They presented a hill-climbing greedy algorithm to solve IM problem and proved that greedy solution is approximated to within a factor of $(1 - \frac{1}{e} - \varepsilon)$.

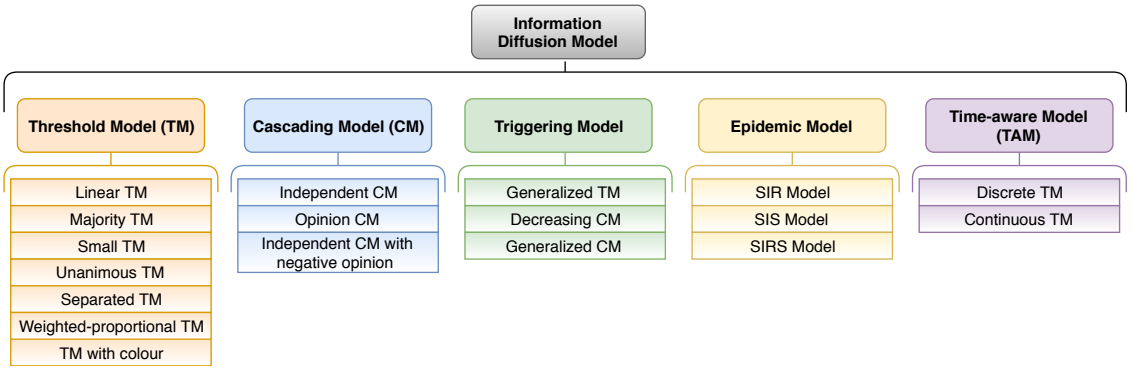

FIGURE 2.1: Taxonomy of Information Diffusion Models [1]

2.1 Information Diffusion Models (IDMs/DMs)

There are huge amount of literature exist in the field of network, epidemiology, database, and data mining, etc. As the necessity of this thesis, we will focus on the diffusion models that are helpful to review

algorithmic aspects of IM approaches. Before discussing diffusion models in detail, we present the generic framework of diffusion models for IM in network. The framework associates each node $u \in V$ belongs to one of the two state: inactive and active at any time-stamp t. Inactive nodes are those who are not influenced by their neighbors or have not heard about the product. Initially, all nodes other than seed nodes are inactive. Seed nodes $S \subseteq V$ are active at time-stamp $t = 0$, i.e., start of the diffusion process. Diffusion process start from seed nodes and these nodes influence their neighbors to be active. Further, the newly active nodes influence their neighbors, and so on. The diffusion process terminates when no new nodes can be activated.

Each model has own mechanism to adopt a new product or thought by capturing active state from inactive based on their neighbors behavior. In this section, we present a brief discussion of models present in taxonomy given in FIGURE 2.1.

- **Threshold Model (TM).** Granovetter and Schelling [59, 60] were the first to introduced threshold models. TM is any model where a threshold value, or a set of threshold values, is used to differentiate ranges of values where the behavior predicted by the model. The most popular threshold diffusion model is linear threshold model. The idea behind this model is that a node y becomes active iff the sum of active neighbors N^a incoming influence must be greater than the threshold

TABLE 2.1: The Comparison of the Characteristics of the Diffusion Models [1]

Diffusion Model	Activation Condition	Properties	Applications
Linear TM [12]	$\sum_{x\in N^a(y)} W_{xy} \geq T_y$	The objective function $\sigma(S)$ is submodular and IM problem is NP-hard under LTM	Rumors and diseases control
Majority TM [53, 54]	$T_y = \frac{1}{2}D(y)$; T_y is threshold value	IM problem is NP-hard under MTM	Distributed computing, Voting system
Small TM [55]	$\sum_{x\in N^a(y)} W_{xy} \geq T_y$; T_y is a small constant	For $T_y = 1$, IM is P-hard and For $T_y \geq 2$ IM is NP-hard under STM	
Unanimous TM [53]	$\sum_{x\in N^a(y)} W_{xy} \geq T_y$; $T_y = d(v)$	2-Approximation algorithm,IM is NP-hard	Network security and vulnerability
Separated TM [12, 42]	$\sum_{x\in N^a(y)\cap\varphi_A^{t-1}} w^A_{x,y} \geq T^A_y$	The objective function is monotonic but not submodular and IM is NP-hard	Network with competitive sources
Weighted-proportional TM [12]	$P[y \in \varphi_A^t \mid y \in \varphi^t \setminus \varphi^{t-1}] = \frac{\sum_{x\in\varphi_A^t} w_{x,y}}{\sum_{x\in\varphi^t} w_{x,y}}$	$\sigma(S)$ is neither submodular nor monotonic and IM is NP-hard.	Deal with two competitive influence
Independent CM [11]	$\prod_{i=1}^{k}(1 - P_y(v_i\|S \cup M_i)) = \prod_{i=1}^{k}(1 - P_y(v_i\|S\cup M_i))$	$\sigma(S)$ is submodular and IM is NP-hard.	Collective behavior, Viral Marketing
Opinion CM [56]	$\sum_{x\in N^a(y)} W_{xy} \geq T_y$	$\sigma(S)$ function is neither submodular nor monotonic and IM is NP-hard	Political campaign and Incorporate user opinions
ICM-NO [57]	$\prod_{i=1}^{k}(1 - P_y(v_i\|S \cup M_i)) = \prod_{i=1}^{k}(1 - P_y(v_i\|S\cup M_i))$	With probability P, each newly active node become positive and with probability 1−P.	Political campaign and Incorporate negative opinions
Decreasing CM [58]	$P_y(x\|S) \leq P_y(x\|M)$	The objective function is submodular and IM is NP-hard under DCM	Collective behavior, Information spreading
SIR [12, 13]	–	$\frac{dS}{dt} = -\beta SI$, $\frac{dI}{dt} = \beta SI - \gamma I$, $\frac{dR}{dt} = \gamma I$	Epidemiology
SIS [12, 13]	–	$\frac{dS}{dt} = -\beta SI + \gamma I$, $\frac{dI}{dt} = \beta SI - \gamma I$	Epidemiology
SIRS [12, 13]	–	$\frac{dS}{dt} = -\beta SI + fI$, $\frac{dI}{dt} = \beta SI - \gamma I$, $\frac{dR}{dt} = \gamma I - fR$	Epidemiology

T_y, i.e., $\sum_{x\in N^a(y)} W_{xy} \geq T_y$. In linear TM, the value of user's threshold follows uniform distribution over [0,1].

There are some other threshold models are exist in the literature, distinguish by their threshold values like majority TM ($T_y = \frac{1}{2}D(y)$) [53, 54], small TM (T_y is a small constant) [55], Separated TM (linear TM with separate competitive cascade) [12, 42], and unanimous TM ($T_y = D(y)$) [53]. There are some generalized threshold models are presented by replacing the activation function of linear TM by an arbitrary function [54, 61]. Bhagat et al. [54] presented a diffusion model named linear threshold with colour that considers user's experience with a product and captures product

adoption rather than influence. Banerjee et al. [61] further extend the linear TM to handle opinion change of users and allow the users to switch back between active and inactive states.

- **Cascading Model (CM).** Inspired by probability theory and interacting particle system [62], dynamic cascade models are introduced for diffusion. The authors of [63, 64] were first to introduce cascade models in the field of marketing. Independent cascade model (ICM) is well-studied and the most popular model in viral marketing [63]. In this model, when a node x becomes active at time t, it has a only chance to activate its inactive neighbors y with activation probability p_{xy} at stamp $t+1$. The activation process of a node can be considered as flipping a coin. If node y becomes active at $t+1$ then it will never be inactive in future. There are some extension of independent cascade model exist in literature like ICM with negative opinion [65], ICM with positive and negative opinion [66].

- **Triggering Model (TRM).** The triggering model is the generalized form of TM and CM, presented by [11]. The authors also proved that triggering model in TM and CM are equivalent. In the triggering model, each node x is associated with a threshold value T_x and a distribution function f_x that maps to a subset of its neighbors S_x with a probability (likelihood that subset can influence node x). This model independently selects a random subset of neighbors in each

instance of diffusion process for user x. There are two generalized triggering model based on diffusion behavior of TM and CM. Kempe et al. [58] presented a more general model than triggering model, named decreasing cascade model. This model redefines the influence probability of a node x from y as $p_y(x, S_x)$, where S_x represents a subset of active neighbors of x. Decreasing CM enforces $p_y(x, S) \geq p_y(x, T)$, $S \subseteq T$ to capture diminishing return property.

- **Epidemic Model (EM).** The epidemic process has had a major impact on transmission of contagious disease, computer virus infections, political campaign, and information propagations such as rumors and news. In epidemic model, the fixed population divided into three classes: susceptible (S), infectious (I), and recovered (R). Kermack and McKendrick [13] presented three epidemic model based on nature of the model cascades: susceptible infectious recovered (SIR), susceptible infectious susceptible (SIS), and susceptible infectious recovered susceptible (SIRS).

- **Time-aware Model (TAM).** The above discussed diffusion models are time-unaware models. These models terminates diffusion process when no more nodes activated. However, some propagation campaigns need to maximize spread under a bounded time. To meet such time-critical demand, time-aware (TAM) models are introduced. The TM is divided into two classes: discrete-time model (DTAM) and continuous-time model (CTAM). The authors of

[57, 67, 68] presented DTAM models by extending independent CM. These models follow discrete random variable over distinct time-stamps. To handle the scenario, when a node x influencing other in continuous time, CTAM [68, 69] were introduced.

TABLE 2.1 summarize each diffusion model. Column Diffusion Model gives the algorithm name with reference. Column Activation Condition describes the condition to change the state from inactive to active. Columns Properties and Applications provide the information of properties and their applications respectively.

2.2 Problem Hardness

In this section, we will discuss the computational hardness of IM problem under TM, CM, TRM, and TAM.

Theorem 2.1. *The influence maximization problem is NP-hard under independent cascade model (ICM) [Theorem 2.4, [11]].*

Theorem 2.2. *The influence maximization problem is NP-hard under linear threshold model (LTM) [Theorem 2.7, [11]].*

Theorem 2.3. *The influence maximization problem is NP-hard under continuous time-aware (CTAM) diffusion model [Theorem 3, [70]].*

Theorem 2.4. *The influence maximization problem is NP-hard under triggering (TRM) diffusion model [11].*

Theorem 2.5. *Computing the expected influence spread $\sigma(S)$ of a seed set S is #P-hard under the independent CM model [Theorem 1, [71]].*

Theorem 2.6. *Computing the expected influence spread $\sigma(S)$ of a seed set S is #P-hard under the LTM model [Theorem 1, [72]].*

Based on observation of above theorems, we can summarize that there is no algorithm exist in literature to find optimal seed set in polynomial time unless $P = NP$. The computation of influence spread $\sigma(S)$ of seed nodes S under any diffusion models is also complex. Therefore, existing research focus on developing approximate and efficient IM algorithms.

2.3 Overview of IM Approaches

In this section, we will review the existing IM algorithms. Although IM problem is NP-hard, the optimal solution can be approximated iff the objective function $\sigma(S)$ is submodular. An arbitrary function is called submodular if it satisfies following two properties.

Property 2.3.1. *(Monotone increasing)* An objective function $\sigma(S)$ is follows monotone increasing iff $\sigma(S) \leq \sigma(T)$, $S \subset T$.

Property 2.3.2. *(Diminishing return)* An objective function $\sigma(S)$ is follows diminishing return iff $\sigma(S \cup u) - \sigma(S) \geq \sigma(T \cup u) - \sigma(T)$, $\forall u \in T$ and $S \subset T$.

The monotonicity states that addition of more nodes in seed set does not reduce its expected influence, while diminishing return means that marginal gain of node u with a subset of seed set is always more or equal to marginal gain with seed set.

Theorem 2.7. *The expected influence spread function $\sigma(S)$ is submodular under ICM [Theorem 2.2, [11]].*

Theorem 2.8. *The expected influence spread function $\sigma(S)$ is submodular under LTM [Theorem 2.5, [11]].*

Theorem 2.9. *The expected influence spread function $\sigma(S)$ is submodular under TRM [Theorem 4.2, [11]].*

Theorem 2.10. *The expected influence spread function $\sigma(S)$ is submodular under CTAM [Theorem 4, [70]].*

2.3.1 The Greedy Framework

Most of the existing work is based on the greedy framework proposed by Kempe et al. [11], which is demonstrated in **Algorithm 1**. The greedy algorithm starts with an empty seed set (line 1) and it iteratively identify a node x with maximum marginal gain (line 3). Then algorithm add node x to seed set S (line 4). Finally algorithm returns the k distinct nodes as resultant seed set.

The theoretical approximation guarantee of the solution generated by greedy algorithm is depends on the objective function $\sigma(S)$ submodular

Algorithm 1: The Greedy framework

Input: Influence graph G, Seed size k.
Output: Seed set S.

$S \leftarrow \phi$
for $i = 1, 2, \ldots, k$ **do**
  $x \leftarrow argmax_{x^* \in V \setminus S}(\sigma(S \cup \{x^*\}) - \sigma(S))$
  $S \leftarrow S \cup \{x\}$
Return S

nature, which holds by the classical IDMs as stated in theorems 2.7 to 2.10.

Theorem 2.11. *Let S^* is the most optimal seed set and S is the seed set returns by the **Algorithm 1**, then we have: $\sigma(S) \geq (1 - (1 - \frac{1}{k})^k)\sigma(S^*)$ [Theorem 2.2, [73]].*

The approximation ratio often simplified as $(1 - 1/e)$ in the literature because of $(1 - \frac{1}{e}) < (1 - (1 - \frac{1}{k})^k)$ for $k > 0$ and $\lim_{k \to \infty}(1 - (1 - \frac{1}{k})^k)$. Moreover, to account the sampling error in sampling algorithms an additional term ε is introduced, i.e., the approximation ratio for sampling methods is $(1 - \frac{1}{e} - \varepsilon)$.

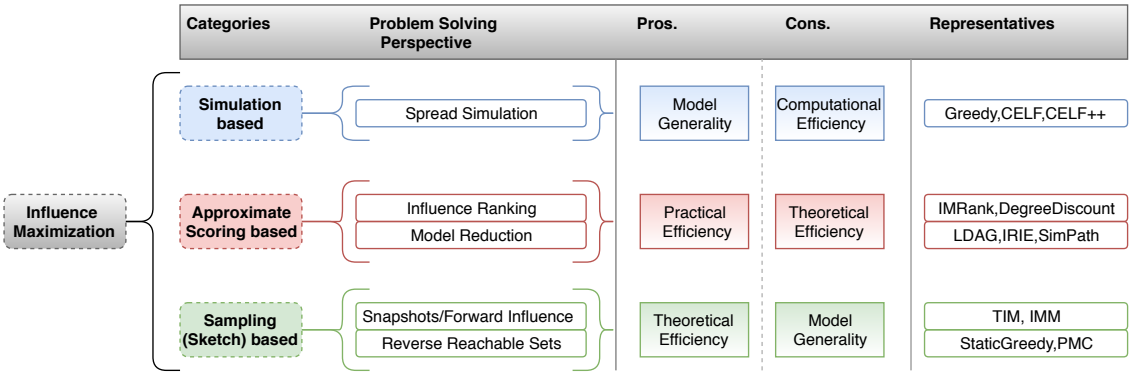

FIGURE 2.2: Taxonomy of the IM Approaches [2, 3]

2.3.2 Taxonomy of Existing IM Approaches

Although greedy framework guaranteed $(1 - \frac{1}{e} - \varepsilon)$ approximation ratio, IM is still challenging to estimate expected influence spread. Theorems 2.5 and 2.6 stated that the evaluation of influence spread $\sigma(S)$ is #P-hard even under simple models. This theoretical hardness leads to an extensive researches on developing efficient approaches to solve IM problem. We divided existing research work into three classes based on how an IM algorithm overcomes the #P-hardness of influence spread estimation: simulation based, approximate scoring based, and sampling based IM approaches. FIGURE 2.2 illustrates the taxonomy of existing IM techniques.

2.3.3 Simulation based IM Approaches

The simulation based IM approaches perform time-consuming Monte-Carlo (MC) simulations to estimate expected influence spread of seed over the network. The seed selection process in these approaches performs r number of explicit MC simulations for each node and estimate average influence of each node. The node with highest average influence is considered as seed and added to the seed set. In each subsequent iterations, nodes those are selected as seed are marked to ignored for computing influence spread. This process iteratively computes seed until k seeds are selected. Svirdenko et al. [27] modified the IM problem which

is defined in [11] by adding node price constraint. In their model k seed nodes are selected with different node prices unlike the traditional IM problem with unit node price.

Due to large number of MC simulations (e.g., $r = 10,000$), the greedy is not efficient for large-scale network. This limitation triggers the researchers to focus on the optimization of the algorithm, which can classified in two classes: reducing number of MC simulations [28, 29, 74] and reducing MC complexity [33, 75].

Minimization of number of MC simulations. There are some algorithms are presented to evaluate an upper bound of $\sigma(S)$ and $\sigma(S \cup \{x\})$ to prune non-candidate seed in subsequent iterations. Leskovec et al. [28] proposed an approach named cost-effective lazy forward (CELF) which is 700 times more efficient than greedy algorithm. CELF uses **Property 2.3.2** of a sub-modular function of cascade influence. In each iteration of the greedy algorithm, it maintains the marginal gain of every node x, given by $\sigma(x|S) = \sigma(x \cup S) - \sigma(S)$. If the marginal gain of a node u at the time $(t+1)$ is more than the other nodes marginal gain at the time t then other nodes marginal gain at the time $(t+1)$ must be less than the marginal gain of node x at the time $(t+1)$. So there is no need to evaluate marginal gain of other nodes at the time $(t+1)$ which significantly improves the time efficiency of the algorithm.

Inspired by CELF, Goyal et al. [29] proposed CELF++ algorithm. In each round of iteration, CELF++ computes marginal gain of node x with

current seed set S and $(S \cup y)$ are given as, $\sigma(x|S) = \sigma(x \cup S) - \sigma(S)$ and $\sigma(x|(S \cup y)) = \sigma((x \cup (S \cup y)) - \sigma(S \cup y)$ respectively. Node y is maximum marginal gain node by now in the current iteration. CELF++ computes two marginal gain values simultaneously, as a result, it is 30% to 50% faster than CELF experimentally. Zhou et al. [74] presents an algorithm to estimate an upper bound of influence spread of each individual using matrix analysis. The matrix analysis approach obtains an upper bound of $\sigma(x)$ by a few multiplications of sparse matrix to avoid initial iteration of CELF/CELF++.

Improving the complexity of MC simulations. The other way to improve efficiency of greedy algorithm is to reduce the search space for each individual MC simulation. Wang et al. [33] proposed an approach named community based greedy algorithm (CGA). To improve MC complexity, CGA first divides the network into subnetworks and then identify influential users within subnetwork using divide-and-conquer approach. It proposed a cost function to identify community structure in mobile networks. The improvement of efficiency is limited and bound to social structure of network. Therefore, CGA is not suitable for large-scale networks.

Chen et al. [49] proposed an algorithm viz. community-based influence maximization (CIM). It uses the community-based framework to identify seed nodes efficiently. CIM first detect community structure of influence graph then it selects the seed nodes from each community based on their

seed quota. Thus it works efficiently, although influence spread is still an issue. The advantage of CIM and CGA is the reduction of search space by partitioning of network, that MC simulations run only on subnetworks.

Sheng et al. [76] presented an algorithm LPIMA based on community detection to identify seed nodes in the network. The CGA algorithm improves the efficiency by reducing search space using partitioning but it needs to simulate marginal gain of each node in the community. Therefore, to improve the efficiency of CGA, they proposed LeaderRank method to simulate influence of community nodes. Then, LPIMA selects candidate nodes. based on quantify value of influence. They also incorporate submodular property to further improve the efficiency of greedy. The experimental results show that LPIMA is more efficient than CGA.

2.3.4 Approximate Scoring based IM Approaches

Approximate scoring based IM approaches evaluate $\sigma(x)$ for each node x by an approximate scoring method unlike simulation based approaches. These approaches avoid the use of time-consuming MC simulations. Therefore, these approaches are more efficient and scalable. Most of scoring based approaches are model specific and account the properties of corresponding model. These models can be categories into two classes:

influence ranking (rank refinement) and model reduction based approaches.

Rank refinement. Rank refinement methods assign the ranking to each individual based on some approximated metrics for estimating influence spread. Then seed nodes are easily selected based on their ranking. There are some simple rank refinement methods exist like, degree, distance centrality [77], and PageRank [78] to select seed users. However, these methods do not provide good quality solution, because of these methods avoid influence overlaps and features of diffusion models. To overcome the weaknesses of simple ranking methods, there are some influence ranking methods like GROUP-PR [79], Degree Discount [24], IRIE [80] etc., are introduced.

Chen et al. [24] proposed an algorithm named Degree Discount. The idea behind this approach is that a node after selecting as seed, no longer available to be influenced by its neighbors. So, degrees of its neighbors are reduced by one. Initially, it selects highest degree node of the network and reduces the degree of its neighbors by one. Iteratively it selects highest degree node and adds it to seed set S, followed by degree discount step. This algorithm outperforms highest degree approach in terms of accuracy. However, this improvement is very limited.

The authors of [79] extend the idea of PageRank from an individual to a set of users, named as GROUP-PR. This method estimates the ranking score of a node set S by adding the PageRank score of each individual

$x \in S$. GROUP-PR is an influence upper bound estimation method and follows greedy framework. First, it estimate the PageRank score of each individual. Then, it iteratively finds the maximum marginal gain node x in term of ranking score and adds x to S. Finally, GROUP-PR returns k-node set as most influential users. GROUP-PR is more efficient because of it avoids native MC simulations and more accurate than PageRank.

Jung et al. [80] proposed a more general PageRank method named as IRIE. This method gives some ranking to each user based on message passing influence estimation. It design a system of $|V|$ equations with $|V|$ varies to compute influence of each individual $x \in V$. To estimate influence spread of x, it combines the influence of itself with its direct influence to neighbors i.e., $(1 + \alpha \sum_{y \in N_{out}(x)} p_{xy}.Inf_y)$, where $\alpha \in [0,1]$. This method solves the system of linear equation iteratively and after k iterations it returns seed set S. IRIE matches influence spread with greedy with high efficiency. There are some other rank refinement algorithms such as SPIN [81], IMRANK [82], etc.

Model reduction. The model reduction methods simplifies the information diffusion process to address the #P-hardness of estimating $\sigma(S)$ of seed S. There are two ways to handle model reduction in IM problem: reduction of stochastic models to deterministic models for estimation of exact spread, and restriction of influence to local region.

Kimura et al. [30] proposed the shortest path based approaches, shortest path model (SPM) and shortest path 1 model (SP1M). The authors

assumed that only shortest and second shortest paths play a role in influence spread. So that influence spread can be computed recursively by the Dijkstra shortest-path algorithm. SP1M does not need MC simulations. An approximation strategy is used to estimate objective function for improving its performance. SPM/SP1M only considers path length and avoids their influence probability. Therefore, it can not establishes a good approximation ratio.

Inspired by SP1M, Chen et al. [31] proposed an algorithm named maximum influence arborescence (MIA). To estimate influence spread, it maintains local arborescence structure. MIA considers only highest propagation probability paths to evaluate influence spread and restricts the influence spread to local tree structure. Therefore, It is more time efficient and avoids MC simulations. The authors of [31] also present a parallel variant of MIA, named as PMIA, which is computationally more efficient. Kim et al. [32] proposed an approach named independent path algorithm (IPA) based on reduction of ICM, similar to MIA. IPA assumes that influence paths from node x to y are independent of each other. It considers all the paths from node x to y whose propagation probability is more than threshold unlike MIA/PMIA. Influence spread of paths is computed parallelly, therefore IPA is more time-efficient.

Using directed acyclic graph (DAG) structure, Chen et al. [83] introduced an algorithm viz. LDAG. It has the similar idea as PMIA, but tailored for LTM. For each node x, it constructs local DAG structure using Dijkstra

shortest-path algorithm and assumes that influence of x is limited to its local region. The construction of LDAG for a node x is both computationally and memory intensive for large-scale networks. It uses greedy hill-climbing approach for seed selection which evaluates seed linearly. Therefore, LDAG is more scalable and tractable.

Goyal et al. [84] introduced a scoring-based IM approach under LTM, named as SIMPATH. It estimates the influence spread of seed nodes or a set of nodes by enumerating all simple paths starting from seed nodes. SIMPATH uses a pruning strategy to restricts small neighborhood. It uses the vertex cover optimization technique to decrease the running time in the first iteration. It uses look ahead optimization in subsequent iterations. It generates the same level of influence spread as greedy. SIMPATH is more time and space efficient than LDAG.

The authors of [43] proposed another proxy method for both ICM and LTM diffusion models, named EASYIM. It enumerates all simple paths within length l to estimates the influence spread of seed nodes. To improve the accuracy of method, it also accounts the overlaps between paths. EASYIM incorporates the IRIE method to estimate global influence in an iterative manner and achieves better output.

2.3.5 Sampling based IM Approaches

Sampling based IM approaches improve theoretical efficiency of simulation based IM approaches with an approximation guarantee. These approaches perform prior computation of a number of *sketches* under specific diffusion model to avoid rerunning of time-consuming MC simulation. These approaches computes influence spread by exploiting the sketches. Sampling based IM approaches are categories into two classes based on how sketches are generated : snapshot based and reverse reachable (RR) set.

Snapshot based sampling. The idea of snapshot based sampling approaches is to construct a sketch or subgraph by extracting an instance of influence diffusion process regarding specific model such as ICM and LTM. Next, it computes the expected influence spread of the seed set S accurately using these snapshots with approximation guarantee. For example, a graph $G = (V, E, W)$ with a diffusion model ICM constructs a snapshot by removing each edge (x, y) with $(1 - p_{x,y})$ probability and form a subgraph G_i. Let $I_S(G_i)$ represents the influence spread of seed S on G_i and $\{G_i, G_2, \ldots, G_m\}$ are the m-snapshots of G at different instances. Then, the expected influence spread of seed set S is the average of influence spread of S on these snapshots, i.e., $\sigma(S) = \frac{1}{m}\sum_{i=1}^{i=m} I_S(G_i)$. The greedy framework with snapshot sampling can achieve $(1 - \frac{1}{e} - \varepsilon)$ approximate solution. Now, we review the snapshot sampling based IM approaches.

Chen et al. [24] proposed a sampling method based on forward influence under ICM diffusion model, named as NewGreedy. It constructs a number of snapshots at different instances to evaluate marginal gain $(\sigma(S \cup x) - \sigma(S))$ of each node $x \in V \setminus S$ at each iteration of NewGreedy. The asymptotic complexity of constructing a snapshot G_i of G is equal to the complexity of running a MC simulation. Therefore, NewGreedy is significantly better than simulation based greedy [11] regarding efficiency with $(1 - \frac{1}{e} - \varepsilon(r))$ approximate influence spread.

Cheng et al. [85] proposed a sampling algorithm viz. StaticGreedy which guarantee the sub-modularity of objective function $\sigma(S)$ in seed selection process. It constructs $(m = (8 + 2\varepsilon)N\frac{\log N + \log\binom{N}{k} + \log 2}{\varepsilon^2})$ snapshots with $(1 - n^{-1})$ probability to achieve $(1 - \frac{1}{e} - \varepsilon)$ approximate solution. StaticGreedy is much faster than simulation greedy. Although worst case time complexity of StaticGreedy is still high. To further improve its efficiency, StaticGreedyDU [85] is introduced. It uses a pruning strategy to empirically reduce its efficiency. StaticGreedyDU pruned each node reachable from seed set S_i at iteration i from all snapshots, and subsequent influence spread estimation performed on pruned snapshots, which would improve its efficiency.

PRUNEDMC [86] is introduced to further improve the time-efficiency of StaticGreedyDU [85] using an index structure. It builds a DAG for each snapshot G_i, and every node x in DAG a strong connected component of snapshot G_i. A maximum degree node is selected from each DAG as hub

node. To built an index structure, it marks ancestors and descendants of hub node on the snapshots. PRUNEDMC speed up the estimation of influence spread of x by avoiding the traversal of descendants of hub node, if x is the ancestor of hub node for corresponding snapshot. Therefore, the running time of marginal gain estimation of a node x is effectively reduced by combing index structure with StaticGreedyDU pruning technique.

Cohen et al. [87] introduced a snapshot sampling approach SKIM. In order to speed up the computation of influence spread on constructed snapshots, this method uses bottom-K^2 min-Hash. It performs reverse Breadth First Search (BFS) on snapshot and updates bottom-K^2 min-Hash values simultaneously for a number of candidate seed sets. SKIM is faster than simulation based algorithms and some heuristic algorithms, but its worst-case complexity is equal to StaticGreedy.

Reverse reachable (RR) sets. Borgs et al. [88] are the first to introduce the concept of reverse reachable sets in IM problem. They state that the estimation of influence spread on sketches constructed by operating on whole graph is not necessary. In the reverse reachable approach, the estimation of influence spread of a seed set S is based on the selection of random nodes and seeing the portion of selected nodes which can be reached by seed set S. Based on RR sets, Borgs et al. introduced a threshold-based method, named as RIS. In this method, They construct RR sets continuously until each edge $(x,y) \in E$ examined during RR sets construction process reaches a threshold θ.

Tang et al. [89] proposed TIM, to make RR approach more practicable and efficient. It improves RIS by better analysis on required number of RR sets to achieve same theoretical approximation bound. TIM requires m RR sets to ensure theoretical bound, where $(m = O(\frac{\varepsilon^{-2}N(\log N + \log \binom{N}{k})}{IS_{optS}}))$ and IS_{optS} denotes the influence spread of optimal seed set. The authors of [89] also introduce another variant of TIM by improving parameter estimation process, known as TIM+. It gives better empirical performance than TIM with same worst-case time complexity. To further improve the performance of TIM/TIM+, Tang et al. [90] proposed a martingale approach, known as IMM. It has better bootstrap parameter estimation procedure than TIM/TIM+. Therefore, IMM is more time-efficient than TIM/TIM+.

From above discussed RR sampling based approaches, i.e., RIS [88], TIM/TIM+ [89], and IMM [90], it is important to notice that these approaches may required a large memory space. This is because of two reasons: a large number of RR sets are generated to preserve theoretical approximation, and every RR set should be stored in the memory for seed selection process by greedy procedure.

In order to overcome memory limitation, the authors of [91] proposed BKRIS method based on lazy sampling approach. First, BKRIS computes lower bound on influence spread of optimal seed set IS_{optS}. Next, it estimates the number of RR sets m using lower bound on IS_{optS}. Similar to SKIM [87], BKRIS adopts bottom-K min-Hash strategy. Then, it

computes seed set with fully utilizing every RR sets unless necessary. The lazy sampling strategy of BKRIS practically speeds up the IMM by two orders of magnitude. To further improve the performance of IMM, the authors of [92] proposed an orthogonal stop-and-stare optimization approach SSA. This approach doubles the number of RR sets iteratively and generate seed sets using current generated RR sets. It stops the iteration whenever estimated influence $\sigma(S_i)$ at iteration i is close to estimated influence $\sigma(S_{i-1})$ computeed at iteration $(i-1)$. They also present an improved variant of SSA named D-SSA. The authors also claim that SSA/D-SSA ensure $(1-\frac{1}{e}-\varepsilon)$ approximation ratio.

2.3.6 Summary and Discussion.

Simulation based approaches. These approaches are developed to improve the efficiency of the greedy algorithm. These approaches uses MC simulations as black-box, i.e., model generality is preserved but prevents performance improvement by utilizing diffusion models properties. SA [75] is an exception among all simulation based approaches, as it does not ensure any approximation guarantee. This is because it uses simulated annealing meta-heuristic to explore and search seed nodes in the network. SA may stuck in local optima as it does not provide theoretical guarantee. This algorithm could perform slightly better than other greedy approaches in term of influence spread with less running time.

TABLE 2.2: The Comparison of the Characteristics of The Existing IM Algorithms – I

Algorithm	Time Complexity	Approximation	Problem Solving Perspective	State-of-the-art Algorithms	Base Algorithm
Greedy [11]	$O(kNMI)$	$1-1/e-\varepsilon$	Spread Simulation	MaxDegree, Central & Random	–
Knapsack Greedy [27]	$O(N^5)$	$1-1/e-\varepsilon$	Spread Simulation	–	Greedy
SP1M [30]	$O(kNM)$	$1-1/e$	Influence Path	Degree, PageRank & Closeness	–
CELF [28]	$O(kNMI)$	$1-1/e-\varepsilon$	Sub-modularity	Greedy	Greedy
Degree Discount [24]	$O(k\log N+M)$	N.A.	Heuristic based	CELF, Greedy & Random	MaxDegree
NewGreedy [24]	$O(kIM)$	$1-1/e-\varepsilon(r)$	Snapshots	CELF, Greedy & Random	MaxDegree
TW Greedy [25]	$O(kNMI)$	$1-1/e-\varepsilon$	Spread Simulation	SCG, KKG & High Degree	Greedy
MIA / PMIA [31]	$O(Nt_{i\theta}+kn_{o\theta}n_{i\theta}(n_{i\theta}+\log N))$	$1-1/e$	Influence Path	Greedy, Random, DD & PageRank	SP1M
LDAG [83]	$O(Nt_\theta+kn_\theta m_\theta(m_\theta+\log N))$	N.A.	Score Estimation	Greedy,SPIN, DD & PageRank	–
CGA [33]	$O(M+IM_C(N(Z-C)+k(C+N_C)))$	$1-e^{-1/(1+\delta_C)}$	Community Based	DD, MG & Random	–
CELF++ [29]	$O(kNMI)$	$1-1/e-\varepsilon$	Sub-modularity	CELF	CELF
SA [75]	$O(TIM)$	N.A	Spread Simulation	NewGreedy, CGA, & DD	Greedy
Diffusion Degree [26]	$O(N+M)$	N.A.	Centrality Based	DD & High Degree	High Degree
SIMPATH [84]	$O(klNP_\theta)$	N.A.	Score Estimation	High Degree, CELF & PageRank	LDAG
IRIE [80]	$O(k(n_{o\theta}k+M))$	N.A.	Score Estimation	Greedy & PMIA	–
BP-Greedy [93]	–	$1-1/e$	Spread Estimation	Betweenness, Degree, & Closeness	Greedy
IPA [32]	$O(\frac{NO_v n_{v\mu}}{c}+k^2(\frac{O_v n_{v\mu}}{c}+(c-1)))$	N.A.	Influence Path	Greedy, DD & Random	PMIA
StaticGreedy [85]	$O(\frac{kMN^2\log\binom{N}{k}}{\varepsilon^2})$	$1-1/e-\varepsilon$	Snapshots	CELF, SP1M, DD & High Degree	PMIA
PRUNEDMC [86]	$O(\frac{kMN^2\log\binom{N}{k}}{\varepsilon^2})$	$1-1/e-\varepsilon$	Snapshots	IRIE, Random, PMIA & Degree	Greedy
TIM [89]	$O(\frac{k(M+N)\log N}{\varepsilon^2})$	$1-1/e-\varepsilon$	Reverse Reachability	CELF++, IRIE & SIMPATH	–
GROUP-PR [79]	$O(kMN)$	N.A.	Influence Ranking	CELF, IRIE,DD & PMIA	PageRank
RIS [88]	$O(\frac{k(N+M)\log^2 N)}{\varepsilon^2}$	$1-1/e-\varepsilon$	Reverse Reachability	–	–
IMRANK [82]	$O(NTd_{max}\log d_{max})$	N.A.	Rank Refinement	PMIA & IRIE	–
SKIM[87]	$O(\frac{kN^2Mlog\binom{N}{k}))}{\varepsilon^2})$	$1-1/e-\varepsilon$	Reverse Reachability	TIM	–
IMM [90]	$O(\frac{(k+l)(N+M)\log N)}{\varepsilon^2})$	$1-1/e-\varepsilon$	Reverse Reachability	TIM, TIM+ & IRIE	–
MPMN-CELF++ [41]	$O(kNMI)$	N.A.	Spread Simulation	SIMPATH & CELF++	CELF++
MPMN-SIMPATH [41]	$O(klNP_\theta)$	N.A.	Influence Ranking	SIMPATH & CELF++	SIMPATH
UBLF [74]	$O(kINM)$	$1-1/e-\varepsilon(I)$	Spread Simulation	Greedy & CELF	Greedy
EASYIM [43]	$O(kD(N+M))$	N.A.	Influence Ranking	SIMPATH, CELF++ & IRIE	Greedy
LCI [38]	$O((N+M)N.d)$	N.A.	Sub-modularity	Greedy	Greedy
SSA/D-SSA [92]	–	N.A.	Reverse Reachability	IMM & TIM+	RIS
ASMTC [39]	$O(\lvert V^s\rvert^2+\lvert V^s\rvert)$	N.A.	Reverse Reachability	–	–
SeedSelection-M [40]	–	N.A.	Rank Refinement	Degree, K-Shell & VoteRank	–
BKRIS [91]	$O(\frac{NM(\log N+\log\binom{N}{k}))}{\varepsilon^2})$	$1-\frac{1}{e}-\varepsilon-\varepsilon'$	Reverse Reachability	RIS	IMM
DPSO [47, 94]	$O(k^2\log knD^2)$	N.A.	Swam Optimization	Degree, CELF++ & SAEDV	–
LAIM [48]	–	$1-1/e-\varepsilon$	Learning Based	Degree, CELF & Random	Greedy

TABLE 2.3: The Comparison of the Characteristics of the Existing IM Algorithms – II

Algorithm	Diffusion Model				Category				Network	
	LT	IC	TRM	CTAM	Simulation	Heuristic	Meta heuristic	Sketch	Single	Multiple
Greedy [11]	✓	✓	✓	✓	✓	✗	✗	✗	✓	✗
Knapsack Greedy [27]	✓	✓	✓	✓	✓	✗	✗	✗	✓	✗
SP1M [30]	✗	✓	✗	✗	✗	✓	✗	✗	✓	✗
CELF [28]	✓	✓	✓	✓	✓	✗	✗	✗	✓	✗
Degree Discount [24]	✓	✓	✓	✓	✗	✓	✗	✗	✓	✗
NewGreedy [24]	✗	✓	✗	✗	✗	✗	✗	✓	✓	✗
TW Greedy [25]	✓	✓	✓	✓	✓	✗	✗	✗	✓	✗
MIA / PMIA [31]	✗	✓	✗	✗	✗	✓	✗	✗	✓	✗
LDAG [83]	✓	✗	✗	✗	✗	✓	✗	✗	✓	✗
CGA [33]	✗	✓	✗	✗	✓	✗	✗	✗	✓	✗
CELF++ [29]	✓	✓	✓	✓	✓	✗	✗	✗	✓	✗
SA [75]	✗	✓	✗	✗	✓	✗	✗	✗	✓	✗
Diffusion Degree [26]	✓	✓	✓	✓	✗	✓	✗	✗	✓	✗
SIMPATH [84]	✓	✗	✗	✗	✗	✓	✗	✗	✓	✗
IRIE [80]	✗	✓	✗	✗	✗	✓	✗	✗	✓	✗
BP-Greedy [93]	✓	✓	✓	✗	✓	✗	✗	✗	✗	✓
IPA [32]	✗	✓	✗	✗	✗	✓	✗	✗	✓	✗
StaticGreedy [85]	✗	✓	✗	✗	✗	✗	✗	✓	✓	✗
PRUNEDMC [86]	✗	✓	✗	✗	✗	✗	✗	✓	✓	✗
TIM [89]	✓	✓	✓	✗	✗	✗	✗	✓	✓	✗
GROUP-PR [79]	✗	✓	✗	✗	✗	✓	✗	✗	✓	✗
RIS [88]	✓	✓	✓	✓	✗	✗	✗	✓	✓	✗
IMRANK [82]	✗	✓	✗	✗	✗	✓	✗	✗	✓	✗
SKIM[87]	✗	✓	✗	✗	✗	✗	✗	✓	✓	✗
IMM [90]	✓	✓	✓	✓	✗	✗	✗	✓	✓	✗
LCI [38]	✓	✓	✗	✗	✓	✗	✗	✗	✗	✓
MPMN-CELF++ [41]	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓
MPMN-SIMPATH [41]	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
UBLF [74]	✓	✓	✗	✗	✓	✗	✗	✗	✓	✗
EASYIM [43]	✓	✓	✗	✗	✗	✓	✗	✗	✓	✗
SSA/D-SSA [92]	✓	✓	✓	✓	✗	✗	✗	✓	✓	✗
ASMTC [39]	✗	✓	✗	✗	✗	✗	✗	✓	✗	✓
SeedSelection-M [40]	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓
BKRIS [91]	✓	✓	✗	✗	✗	✗	✗	✓	✓	✗
DPSO [47, 94]	✓	✓	✓	✓	✗	✗	✓	✗	✓	✗
LAIM [48]	✓	✓	✓	✓	✓	✗	✗	✗	✓	✗

Scoring based approaches. To avoid time-consuming MC simulations, these approaches performs scoring procedure based on rank refinement and model reduction. The rank refinement approaches estimates influence

spread efficiently by transforming IM problem to some easier problems, like PageRank [78], GROUP-PR [79], etc. These transformed problems may not be related to IM, although they are computationally efficient. However, these approaches ignore the diffusion model properties in ranking process. Therefore, some model reduction heuristics are introduced to account properties of diffusion models. Model reduction heuristics are directly inherited from classical diffusion models (ICM, LTM) and use these models properties to compute influence spread of seed nodes. These approaches achieve approximate influence spread as simulation based IM algorithms in most cases. However, these approaches cannot maintain a trade-off between influence spread and efficiency when the influence range of nodes and number of influence paths are large. In addition, these approaches are not model generic, i.e., cannot generalized to other models.

Sampling based approaches. To avoid rerunning of MC simulations, these approaches performs sampling in the graph based on snapshots (forward influence) and reverse reachable sets (backward influence). Most of the snapshot based approaches are presented for ICM, although they can be applied and extended to other models like LTM,TRM, and CATM. This is because these diffusion models are node-independent. These snapshot based methods perform significant better over simulation based approaches regarding efficiency with approximation guarantee. However, the theoretical complexity is still an issue in large-scale network. In

general, the backward influence approaches are much faster than forward influence approaches. This is because of snapshots are developed by examining the whole graph while reverse reachable sets are constructed by visiting only those nodes who can activates random sampled nodes. Hence, theoretical complexity of backward influence methods are significantly better than snapshot based approaches.

TABLE 2.2 and TABLE 2.3 summarize the characteristics of the existing IM algorithms. TABLE 2.2 presents the theoretical analysis such as time-complexity, approximation ratio, problem solving perspective, state-of-the-art and base algorithms, etc., of existing IM approaches. TABLE 2.3 discusses the category, diffusion models and type of the network of the corresponding algorithm.

2.4 Context-aware Influence Maximization

The studies of context-aware IM approaches are emerging in recent years. These approaches are the extensions of the conventional IM problem by further considering some contextual features like location of users, topic of information or product category, time of information diffusion, competitive marketing, and dynamic nature of online social networks, etc. The emergence of context-aware IM supports real-life applications such as viral marketing, epidemiology, election campaign, counter-terrorism efforts, rumor control, tread analysis and sales prediction, social

recommendation, blogosphere, network monitoring, and revenue maximization, more effectively. The taxonomy of context-aware IM approaches is illustrated in FIGURE 2.3.

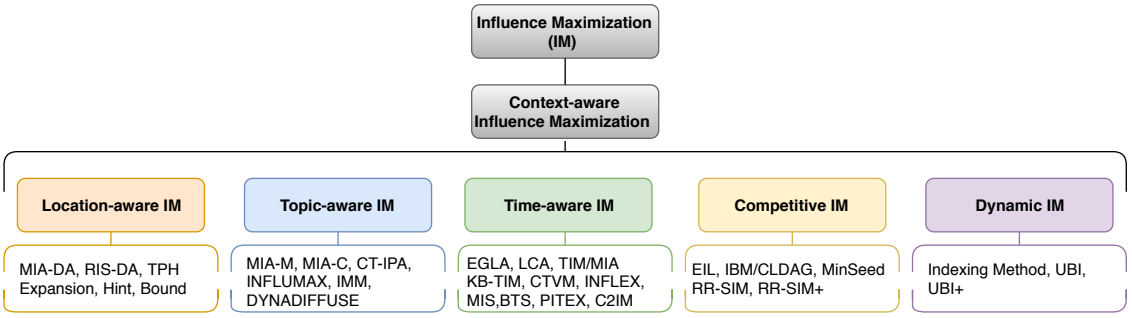

FIGURE 2.3: Taxonomy of the Context-aware IM Approaches [3]

2.4.1 Location-aware Influence Maximization

Location-aware influence maximization (LAIM) problem is the extension of classical IM problem by considering *spatial index* due to commonness of location-based social networks such as Foursquare, Twitter, etc. The objective of LAIM is to maximize the estimated influence spread of location-relevant users unlike generic IM. There are some efforts have been made to solve LAIM problem [95–99].

The authors of [95] are first to introduce LAIM framework, focusing on region queries. Let given a geographical region R, then LAIM is the problem of identifying k users as seed nodes S to maximize the influence spread over region R. They incorporate influence estimation model used in MIA/PMIA [31] under ICM diffusion model. The authors propose an algorithm Expansion which uses best-first search framework to identify

seed. This search procedure accesses users with large upper bounds on influence and prune users with insignificant influence. They also focus on developing upper bounds for pruning. *QuadTree* structure is developed for fast allocation of users with *spatial index*. This work also propose another algorithm Hint, which performs pre-computation of seed S_i for every *QuadTree* leaf regions. Then, combines all S_i as hints to compute upper and lower bound of influence. They perform experiments on real-world location-based social networks to evaluate the performance of LAIM algorithms, and reports running time efficiency in milliseconds for different size regions.

Similar to Expansion/Hint, Wang et. al [98] introduced a distance-aware pruning-based algorithm MIA-DA, which considers user's distance to a query location as edge weights. The authors adopts MIA/PMIA [31] framework under ICM diffusion model to compute influence spread. MIA-DA selects a set of locations as *anchor-locations* to estimate influence bounds. Each anchor-location is selected as a query to compute influence spread. It can utilize triangular equality to compute influence bound for every anchor-location. MIA-DA can also bound with region-based bound estimation in Expansion/Hint. The authors of [97] also adopts distance-aware query and propose an algorithm Target-IM/Target-IM+. They incorporate RR sets instead of MIA/PMIA proxy in [98]. The algorithm develops a pool of tree structures based on weighted RR sets. They also show that the approximation ratio of the

algorithm is $(1 - \frac{1}{e} - \varepsilon)$. The authors of [96] present an algorithm TPH based on distance-aware weighting model.

The authors of [16] proposes IM over trajectory database. They redefines the IM problem as selection of k trajectories on a given advertisement in order to maximize influence spread over a large group of audience. The authors uses a community-based framework that divides the trajectory database into communities in order to find promising trajectories. This work is different from classical IM approaches as it does not incorporate any diffusion models and influence propagation.

The authors of [100] and [101] propose a holistic influence diffusion model (HIM) under ICM settings to spread influence over *spatial* social network. They are first to study how a user x propagate his influence to another y by spatial interactions together with social influence. They first compute RR sets based on *keyword* query Q. Then they present a baseline method SimRmNN to find seed nodes via their spatial interactions. The authors also propose two efficient algorithms Index-based and SimRmNN-Upper to answer HIM queries. These improved algorithms are one or two order of magnitude faster than baseline method.

2.4.2 Topic-aware Influence Maximization

The classical IM problem can be extended by considering *topics* of product being propagated, known as Topic-aware influence maximization

(TAIM) problem. In order to formalize the definition of TAIM, *topic* denotes user's interests as well as product characteristics. The influence spread is dependent on both seed nodes and topics. Then, TAIM is the problem of identifying the optimal seed nodes that maximize the influence over network on given topic-aware query. This problem can be classify into two classes based on users and their relationship characteristics: topic relevant targets and topic dependent diffusion.

Topic relevant targets TAIM. The topic relevant targets TAIM approaches focus on the user's characteristics, i.e., user is topic-aware. There are some efforts such as LGA/ELGA [102], KB-TIM [103], IMIP/IMAX [104], CTVM [105], etc., made on maximizing influence spread over topic-aware users. These approaches differentiate users in the network and compute influence spread $\sigma(S)$ of seed nodes over targeted users.

Li et. al [103] propose a topic-aware query model to identify targeted users based on their profile and estimate seed nodes over targeted space. The user's profile consists the information about the preference of users on specific topics. For example, a user profile $\{< sports, 0.6 >, < tech, 0.5 > , < drama, 0.3 >\}$ denotes the probabilities of user likes to various topics. Then, the algorithm selects targeted users based on a given topic query. The authors incorporate RR sampling strategy and ICM diffusion model to address targeted IM problem. In order to find an unbiased estimator for estimating targeted influence spread, they introduce a weighted sampling

strategy with RR concept. The algorithm select RR sets based on targeted topic query and merges these sets to calculate the result. In order to further reduce the I/O cost, the authors presents an incremental index structure.

The authors of [105] generalize TAIM problem by pre-defined targeted influence function. They adopts RR sampling framework like [103]. The algorithm avoids generation of too many RR sets by an early termination rule. Moreover, they also present the cost-aware setting where activation of an user is dependent on a cost function.

Guo et. al [102] presented a special case of TAIM problem, known as personalized IM problem. This problem identifies a set of seed users to maximize overall influence spread on a given target user which is considered as topic relevant user among all. They adopt ICM and introduce two algorithms *local greedy* (LGA) and *efficient local greedy* (ELGA). LGA is a simulation-based approach with some pruning rules tailored for target user local structure. This approach can not applicable in online query requirement. ELGA is a heuristic approach which considers only shortest path for information diffusion form each node to target node. Therefore, the approximation of influence spread can not be guaranteed theoretically.

Topic dependent diffusion TAIM. The topic dependent diffusion TAIM approaches focus on the user-to-user topic-aware diffusion, i.e., user's relationship with others. Some studies such as AIR-Greedy [15], INFLEX [106], TIM/MIA [107], MIS/BTS [108], etc., focus on topic dependent

diffusion TAIM. These approaches consider the idea of that each edge (x,y) between a pair of individuals x and y is topic dependent. The reason behind this is y might be activated by x on some topics (e.g., tech, drama) and still inactive for others (e.g., sports). To formalize the model, let each edge (x,y) is associated with a propagation probability vector $P_{x,y} = \{pp_1, pp_2, \ldots, pp_t\}$ over t topics and a topic dependent query $Q = \{q_1, q_2, \ldots, q_t\}$. In order to compute $P_{x,y}$ of an edge (x,y) under ICM, it calculates dot product, i.e., $P_{x,y} = \sum_{i=1}^{i=t} q_i.pp_i$.

The authors of [15, 106] were the first to introduce topic dependent TAIM problem. The idea is that similar topic-aware queries will have approximate influence spread for queries. The authors introduced an indexing scheme named as INFLEX based on *similarity-search* and *pre-computation of seed set*. This approach cautiously samples reasonable number of topic distributions queries, and pre-computes seed nodes for each query by any IM approach. Similarly, this approach pre-computes level-2 seed nodes under each query distribution, and combines these level-2 seed sets using a *rank-aggregation* method. Finally, it presents maximum-likelihood Dirichlet estimator, Bregman-ball tree, and Kendell scheme for query sampling, similarity search, and seed set aggregation respectively. The authors of [108] proposed a TAIM approach for some special graph by incorporating similar framework of [106]. These special graphs follow some properties like sub-additive, typically-separable, etc.

In order to provide approximation guarantee, Chen et. al [107] improved

the prior works [15, 106, 108]. The authors adopt MIA/PMIA framework under ICM model. The approximation ratio of the algorithm is bounded under MIA/PMIA framework. To estimate an upper bound of influence spread $\sigma(x)$ of a user x, they present a *best-effort* framework. Then, algorithm estimates exact influence spread and prunes insignificant users based on upper bound influence. The authors also present a topical-sample algorithm to pre-compute seed sets for offline-sampled topic distributions. This algorithm outperforms [15, 106, 108] regarding influence spread with comparable efficiency.

2.4.3 Time-aware Influence Maximization

The classical IM problem terminates the diffusion process when there are no more nodes are adopted the product, idea, or innovation. This condition is practically inefficient and unreasonable as influence propagation process may takes long time. For example, discrete time diffusion models may take $O(N)$ steps and continuous time diffusion models may take an arbitrary time length. Therefore, time-aware influence maximization (TimeAIM) problem introduced a time constraint with propagation model to handle above issue.

Discrete TimeAIM approaches. First, we study some discrete TimeAIM approaches such as MIA-M/MIA-C [67], CT-IPA [68], MISP [109]. Chen et. al [67] introduce the IC-M model, which adopts traditional ICM model.

In IC-M model, a node x contacts y with a meeting probability $m(x,y)$ through edge (x,y). Let node x succeed in meeting with y through edge (x,y) then x has only chance to activate y with $p_{x,y}$ and y has many chances to contact x. TimeAIM under IC-M model selects seed nodes to maximize expected adoption within τ time-steps over random processes. The authors of [68] and [109] presented two identical models CT-IC and LAIC. Under both models, a node x activates y with probability $p(x,y).p_x^{lat}(\Delta_t)$ at $t+\Delta_t$ through edge (x,y), if x is already activated at time-step t. These models are the extension of traditional ICM. In order to find IM solution, they adopt MIA approach. Therefore, these studies do not have theoretical guarantee due to MIA heuristic nature.

Continuous TimeAIM approaches. Rodriguez et al. [69] were the first to introduce continuous-time independent cascade (CTIC) model. The objective function $\sigma(S)$ is monotone and submodular under CTIC model. The authors of [70] proposed an algorithm INFLUMAX which adopts greedy framework for IM with lazy forward optimization. They also described CTIC as continuous Markov chain model (CTMC). The authors of [18, 110] proposed a snapshot sampling based approach to efficiently estimate influence spread with an approximation guarantee. It adopts sampling strategy from SKIM [87].

Xie et. al [111] introduced a new model DynaDiffuse where edge probabilities deteriorate over time. They incorporate CELF greedy [28] and present an optimized greedy approach to compute seed nodes. This

approach does not provide an error bound for stochastic procedure under CTMC diffusion model. Therefore, the algorithm does not provide any theoretical guarantee. Ohsaka et. al [112] proposed a more general time-aware diffusion model time-varying independent cascade (TV-IC). Let a node x activated at time-step t_1 and contact with y through edge (x,y), influence reaches node y at time-step t_2. Then, conditional likelihood of y activating at t_2 is dependent on $(t_2 - t_1)$. The authors also presented first time-aware extension of LTM and names as *time-varying linear threshold* (TV-LT). They proposed a RR sampling based approach for IM under both models TV-IC and TV-LT. The algorithm provides a provable theoretical guarantee due to submodularity of both models.

2.4.4 Competitive Influence Maximization

The competitive influence maximization problem is the extension of generic IM problem by considering competitiveness of products. This problem considers a scenario where multiple competitors follow same social network to spread influence simultaneously. Therefore, the competitive IM aims to maximize own influence spread or minimize opponent influence spread under competitive framework. The existing works can be classified into three classes: known, unknown, and comparative.

IM under known competitor. Carnes et. al [35] and Bharathi [36] were first to introduce competitive IM with known competitor. They consider a scenario where n competitive marketers are present for diffusion over a network and n^{th} marketer want to maximize the spread of influence through diffusion process when seed sets for remaining $(n-1)$ competitors are known. Once node x adopts a product by a marketer then it can not adopts other competitive products. They proved that greedy approach under this framework approximation is same, i.e., $(1-\frac{1}{e}-\varepsilon)$.

The authors of [42] presented various extensions of LTM under competitive framework. They show that the objective function $\sigma(S)$ is non-submodular under these models. The finding of seed nodes which have influence spread more than $\sqrt{\sigma(OPT)}$ is NP-hard, where OPT is optimal seed set under these models. There is another variant of competitive IM problem as *influence blocking maximization* (IBM) is introduced. The idea of IBM is that select a seed set S_2 to minimize the influence spread $\sigma(S_1|S_2)$ of given seed set S_1 for *misinformation* diffusion. The authors of [113] and [114] presented greedy algorithms for IBM problem under ICM and LTM diffusion models respectively. Zhu et. al [115] presented an extension of ICM by considering that a node x can be part of multiple seed sets. They achieve guaranteed approximation $(1-\frac{1}{e}-\varepsilon)$ by adopting greedy framework.

IM under unknown competitor. These approaches consider the scenario that no competitor knows the opponent strategy of seed selection, i.e.

does not have pre-knowledge of opponent seed nodes. The authors of [23] presented competitive IM with unknown competitors as multi-round multi-party game. Li et. al [116] propose a model for competitive IM problem using game theory concept. They model IM problem as Nash equilibrium strategy with n-strategies on a given graph under a diffusion model. Each player cam maximize his own spread using a Nash equilibrium strategy. The classical IM approaches are used as building blocks to find seed solutions of these aforementioned problems.

IM under comparative framework. The diffusion process of comparative IM problem can be classified into two categories: Competitiveness and complementary. In competitive diffusion model, a node x less-likely adopts product P_2, when it already adopts product P_1. Complementary process is voice-versa of competitive diffusion model, i.e., if x adopts P_1 then it has higher chance to adopts P_2. The authors of [117] presented comparative IC model (Com-IC) by extending ICM to tackle comparative framework. Then, they introduced two problems self influence maximization (SIM) and complimentary influence maximization (CIM) to address competitive and complementary diffusion process respectively. The authors extend [89] based on RR sampling to find seed solutions of these problems. Ou et al. [118] also study the competitive IM problem and proposed *interactive* LT model by extending LTM. They propose a heuristic approach TOPBOSS to solve competitive IM problem.

2.4.5 Dynamic Influence Maximization

So far discussed IM approaches consider a static scenario, i.e., social graph $G(V,E,W)$ and diffusion probability are fixed. However, social networks continuously evolving in real-world, i.e. a new users may arrived in the network, a new relationship may be formed. This continuous network formation may affects influence process. Therefore, some efforts are made in dynamic influence maximization (DIM) problem such as MaxG [119], IGA [120], IndexingMethod [121], UBI/UBI+ [122], DIM [123], A-Greedy/H-Greedy [124], etc. DIM approaches can be classified into following categories.

Probing-based DIM strategies. Aggarwal et al. [125] proposed a DIM approach by considering a social graph G and evolution samples of graph over period $[t,t+\theta]$. They proposed a proxy method to identify a seed set S at time-step t such that $\sigma(S|(t+\theta))$ is maximized. It is a simple proxy method which is not align with specific diffusion models. Zhuang et al. [119] proposed a probing-based strategy MaxG to solve DIM problem. They consider that network evolution is detected by *periodically* probing a subset of nodes. This algorithm follows two-phase procedure. In first phase, algorithm selects a set of nodes S_P for probing at time-step $t_s \in [t,t+\theta]$ and constructs a subgraph G_{t_s} by probing S_P. In second phase, it selects a seed set S_{t_s} on G_{t_s} using Degree Discount [24] algorithm to maximize the influence spread of S_{t_s}. This algorithm does not align with any specific diffusion models.

Sampling-based DIM strategies. There are some existing works [122, 126] focus on modeling dynamics of network as snapshots of network $\{G_1, G_2, \ldots, G_T\}$. These DIM approaches continuously select seed sets for each snapshot. Song et al. [122] adopts SP1M [30] and UBLF [74] to estimate influence spread and present a heuristic named upper bound interchange (UBI). Initially, UBI selects seed nodes for G_1 based on offline strategy and estimate influence spread of each node. Then it iteratively updates expected spread of each node. It also updates seed set iteratively for each snapshot by interchanging a seed node by a new node which has influence gain more than 1% of total influence. UBI has no theoretical approximation guarantee due to lack of accurate influence computation and interchange threshold.

The authors of [121] propose an indexing method for DIM problem under IC model. This approach uses RR sets instead of snapshots to consider evolving graph. First, it selects RR sets from G and constructs an indexing structure on these RR sets. Then, it performs re-sampling by deleting and adding nodes from RR sets based on two basic operations SHRINK and EXPAND respectively. Next, it performs set maintenance by sampling any edge or node from a set randomly, and recomputes sample size to manage these samples. Finally, the algorithm selects seed sets from these dynamic maintained RR sets like TIM. This algorithm can be align with other diffusion models like LTM, TRM, etc.

Meng et al. [127] proposed a snapshot sampling-based approach

$T \times oneHope$ under DIM settings. This approach considers T snapshots $\{G_1, G_2, \ldots, G_T\}$ over period $[t_0, t_0 + \Delta t]$ for network evolution, where $T = \Delta t / \omega$. They assume that nodes in each graph G_i are unchanged but edges can be added or deleted over time. They incorporated dynamic ICM model for information diffusion. Finally, they proposed a hop-based approach including recursive formula of activation probability for IM problem under dynamic settings. They pointed that the number of identical users in seed sets are increases if the consecutive snapshots are more similar.

Other DIM strategies. Some works [124, 132] accounts different dynamics of networks like uncertainty and incompleteness of diffusion process. The authors of [132] considers a scenario when propagation probabilities are not given in advance and can be estimated after trails. They present a learning method to acquire propagation probabilities along with diffusion process. They adopt ExploreExploit methods for maximizing influence spread under DIM settings. Tong et al. [124] proposed an adoptive greedy approach for DIM problem by considering propagation probabilities as random variables align with a distribution. The authors of [133] study IM problem over stream data. They use sliding window model to define users influence. In order to continuously detect seed set, they proposed stream influence maximization query.

TABLE 2.4 and 2.5 summarize the characteristics of existing context-aware IM algorithms. TABLE 2.4 provides the contextual

TABLE 2.4: The Comparison of the Characteristics of the Existing Context-aware IM Algorithms – I

Categories	Algorithm	Approximation	Problem Solving Perspective	State-of-the-art Algorithms	Base Algorithm
Location-aware IM	Expansion/Hint [95]	$\varepsilon(1-1/e)$	Model Reduction Heuristic	PMIA,IRIE,Assembly,Bound	–
	TPH [96]	N.A	Rank Refinement	MaxDegree,DD,PageRank	–
	Target-IM/IM+ [97]	$1-1/e-\varepsilon$	Reverse Reachable Sets	MIA-L,Expansion,IMM	MIA
	MIA-DA/RIS-DA [98]	$(1-1/e)$	Model Reduction Heuristic	PMIA,MIA/RIS-DA	RIS
	TOA/TORA [99]	$(1-1/e)$	Model Reduction Heuristic	DD,LIA,CELF,PMIA	WRIS
Topic-aware IM	AIR-Greedy [15]	$1-1/e-\phi$	Spread Simulation	TIC	Greedy
	LGA/ELGA [102]	N.A	Model Reduction Heuristic	LDegree,LRandom,LND	–
	KB-TIM [103]	$1-1/e-\varepsilon$	Reverse Reachable Sets	RR,IRR,WRIS	TIM
	INFLEX [106]	N.A	Rank Refinement	exactKNN,approxKNN,approxAD	–
	TIM/MIA [107]	$\varepsilon(1-1/e)$	Model Reduction Heuristic	PMIA,INFLEX,MIS	MIA
	IMIP/IMAX [104]	$1-1/e$	Spread Simulation	CELF,PMIA,IRIE,CD	–
	MIS/BTS [108]	N.A	Model Reduction Heuristic	TA-PMIA/Greedy/PageRank	PMIA
	CTVM [105]	$1-1/e-\varepsilon$	Reverse Reachable Sets	TIM,CELF++,SIMPATH	BIM
	PITEX [128]	$(1-\varepsilon)/(1+\varepsilon)$	Reverse Reachable Sets	RR,MC,INDEXEST	TIM
	In-out Discounting [129]	N.A	Spread Simulation	In-Out, MaxDegee, DD	–
Time-aware IM	MIA-M/MIA-C [67]	$(1-1/e)$	Model Reduction Heuristic	Greedy,MaxDegree	MIA
	CT-IPA [68]	$(1-1/e)$	Model Reduction Heuristic	Greedy,MaxDegree,Random,IPA	IPA
	MISP [109]	$(1-1/e)$	Model Reduction Heuristic	Random,DC,PMA,ISP	MC
	INFLUMAX [70]	$(1-1/e)$	Markov Chain	Greedy,PMIA,Random,SP1M	–
	FASTMARGIN [111]	N.A	Markov Chain	CELF/FastMargin-Static/Dynamic	CELF
	IMM [90]	$1-1/e-\varepsilon$	Reverse Reachable Sets	TIM,TIM+,SIMPATH	TIM
	TSDEG/TSREEDY [130]	$1-1/e$	Rank Refinement & Simulation	Greedy,Random,BET,MaxDegree	Greedy
Competitive IM	EIL [113]	$1-1/e-\varepsilon$	Spread Simulation	Greedy,Degree,EarlyInfectees	Greedy
	IBM/CLDAG [114]	N.A	Model Reduction Heuristic	Greedy,Degree,Random	LDAG
	RR-SIM/RR-SIM+ [117]	$\alpha(1-1/e-\varepsilon)$	Reverse Reachable Sets	PageRank,MaxDegree,Random	–
	MinSeed [115]	$1-1/e-\varepsilon$	Spread Simulation	LS-Greedy	Greedy
	CI2 [131]	N.A	Snapshot Sampling	Greedy,MaxDegree,INCIM,IPA	INCIM
Dynamic IM	MaxG [119]	N.A	Rank Refinement	Rand,Deg,Enum,DegRR	Probing
	IGA [120]	N.A	Spread Simulation	Random,MaxDegee,HT	Greedy
	IndexingMethod [121]	$1-1/e-\varepsilon$	Reverse Reachable Sets	TIM,IMM,PMC,IRIE	RIS
	UBI/UBI+ [122]	N.A	Rank Refinement	MaxDegree,IMM,IRIE	Greedy
	DIM/Opt-DIM [123]	N.A	Model Reduction	LDAG,SIMPATH	–
	A-Greedy/H-Greedy [124]	N.A	Spread Simulation	Random, Greedy	Greedy

category, name of the algorithm, approximation ratio, problem solving perspective, base and the state-of-the-art algorithms. TABLE 2.5 provides the algorithmic-category, diffusion models, and type of the network where the corresponding algorithm can be applied.

TABLE 2.5: The Comparison of the Characteristics of the Existing Context-aware IM Algorithms – II

Algorithm	Diffusion Model (IDM)			Category			Network			
	ICM	LTM	TM	Simulation	Heuristic	Sketch	Single	Multiple	Static	Dynamic
Expansion/Hint [95]	✓	✗	✗	✗	✓	✗	✓	✗	✓	✗
TPH [96]	✓	✗	✗	✗	✓	✗	✓	✗	✓	✗
Target-IM/IM+ [97]	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗
MIA-DA/RIS-DA [98]	✓	✗	✗	✗	✓	✗	✓	✗	✓	✗
TOA/TORA [99]	✓	✗	✗	✗	✓	✗	✓	✗	✓	✗
AIR-Greedy [15]	✓	✓	✗	✓	✗	✗	✓	✗	✓	✗
LGA/ELGA [102]	✓	✗	✗	✗	✓	✗	✓	✗	✓	✗
KB-TIM [103]	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗
INFLEX [106]	✓	✗	✗	✗	✓	✗	✓	✗	✓	✗
TIM/MIA [107]	✓	✗	✗	✗	✓	✗	✓	✗	✓	✗
IMIP/IMAX [104]	Expectation IDM			✓	✗	✗	✓	✗	✓	✗
MIS/BTS [108]	✓	✗	✗	✗	✓	✗	✓	✗	✓	✗
CTVM [105]	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗
PITEX [128]	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗
In-out Discounting [129]	Target Adoption IDM			✓	✗	✗	✓	✗	✓	✗
MIA-M/MIA-C [67]	✓	✗	✗	✗	✓	✗	✓	✗	✓	✗
CT-IPA [68]	✓	✗	✗	✗	✓	✗	✓	✗	✓	✗
MISP [109]	✓	✗	✗	✗	✓	✗	✓	✗	✓	✗
INFLUMAX [70]	✗	✗	✓	✗	✓	✗	✓	✗	✓	✗
FASTMARGIN [111]	DynaDiffuse IDM			✗	✓	✗	✓	✗	✓	✗
IMM [90]	✗	✗	✓	✗	✗	✓	✓	✗	✓	✗
TSDREEDY [130]	✓	✓	✗	✓	✗	✗	✓	✗	✓	✗
EIL [113]	✓	✗	✗	✓	✗	✗	✓	✗	✓	✗
IBM/CLDAG [114]	✗	✓	✗	✗	✓	✗	✓	✗	✓	✗
RR-SIM/RR-SIM+ [117]	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗
MinSeed [115]	✓	✗	✗	✓	✗	✗	✓	✗	✓	✗
CI2 [131]	Decidable Competitive DM			✗	✗	✓	✓	✗	✓	✗
MaxG [119]	✗	✗	✓	✗	✓	✗	✓	✗	✗	✓
IGA [120]	✓	✓	✗	✓	✗	✗	✓	✗	✗	✓
IndexingMethod [121]	✓	✗	✗	✗	✗	✓	✓	✗	✗	✓
UBI/UBI+ [122]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✓
DIM/Opt-DIM [123]	✗	✓	✗	✗	✓	✗	✓	✗	✗	✓
A-Greedy/H-Greedy [124]	✓	✗	✗	✓	✗	✗	✓	✗	✗	✓

2.5 Performance Metrics

In this section, we explain the performance metrics used in the evaluation of the IM algorithm. There are four major performance metrics present in the literature: quality, efficiency, scalability, and robustness.

TABLE 2.6: The Comparison of the Performance of the Existing IM Algorithms

	Category	Algorithm	Time efficiency	Seed Quality	Memory Footprint	Robustness
Classical IM approaches	Simulation-based	Greedy [11]	✗	✓	✓	✗
		Knapsack Greedy [27]	✗	✓	✓	✗
		CELF [28]	✗	✓	✓	✗
		CGA [33]	✓	✓	✓	✗
		CELF++ [29]	✗	✓	✓	✗
		SA [75]	✓	✓	✓	✗
		UBLF [74]	✓	✓	✓	✗
		LCI [38]	✗	✓	✗	✗
	Scoring-based	SP1M [30]	✓	✗	✓	✗
		Degree Discount [24]	✓	✓	✓	✗
		TW Greedy [25]	✗	✓	✓	✗
		MIA / PMIA [31]	✓	✗	✓	✗
		LDAG [83]	✓	✓	✗	✗
		Diffusion Degree [26]	✓	✓	✓	✗
		SIMPATH [84]	✓	✓	✓	✗
		IRIE [80]	✗	✗	✓	✓
		IPA [32]	✓	✗	✓	✗
		IMRANK [82]	✓	✓	✓	✗
		EASYIM [43]	✗	✓	✓	✓
	Sampling-based	IMM [90]	✗	✓	✗	✗
		StaticGreedy [85]	✗	✓	✗	✗
		PRUNEDMC [86]	✗	✓	✗	✗
		TIM [89]	✗	✓	✗	✗
		RIS [88]	✗	✓	✗	✗
		SKIM[87]	✓	✓	✗	✗
		SSA/D-SSA [92]	✓	✓	✗	✗
		BKRIS [91]	✓	✓	✗	✗
Context-aware approaches	Location-aware	Expansion/Hint [95]	✓	✓	✓	✗
		TPH [96]	✓	✓	✓	✗
		Target-IM [97]	✓	✓	✗	✗
		MIA-DA/RIS-DA [98]	✓	✓	✗	✗
		TOA/TORA [99]	✓	✓	✓	✗
	Topic-aware	AIR-Greedy [15]	✗	✓	✓	✗
		LGA/ELGA [102]	✓	✓	✗	✗
		KB-TIM [103]	✓	✓	✗	✗
		INFLEX [106]	✓	✓	✓	✗
		TIM/MIA [107]	✓	✓	✓	✗
		CTVM [105]	✓	✓	✗	✗
		PITEX [128]	✓	✓	✗	✗
	Time-aware	MIA-M/MIA-C [67]	✓	✓	✗	✗
		CT-IPA [68]	✓	✓	✗	✗
		INFLUMAX [70]	✓	✓	✓	✗
		FASTMARGIN [111]	✓	✓	✓	✗
		IMM [90]	✓	✓	✗	✗
		TSDEG [130]	✓	✓	✓	✗
	Competitive	EIL [113]	✗	✓	✓	✗
		IBM/CLDAG [114]	✓	✓	✗	✗
		RR-SIM/SIM+ [117]	✓	✓	✗	✗
		MinSeed [115]	✗	✓	✓	✗
		CI2 [131]	✓	✓	✗	✗
	Dynamic	MaxG [119]	✓	✓	✓	✗
		IGA [120]	✗	✓	✓	✗
		IndexingMethod [121]	✓	✓	✗	✗
		UBI/UBI+ [122]	✓	✓	✓	✗
		DIM/Opt-DIM [123]	✓	✗	✓	✗

1. **Quality: Influence Spread.** Quality, in IM problem, equates the number of product adoption in network by algorithm, with given seed set S, $|S| = k$. In General, the influence spread grows with k although few minor fluctuations are possible. Recently, some context-aware IM algorithms [15, 99, 104, 120, 124, 129, 131] are introduced to improve the quality/effectiveness of seed.

2. **Efficiency: Running Time.** Efficiency, in IM problem, measured in terms of running time. Efficiency is the ability of an algorithm to produce a desired result, i.e., seed set S in efficient time. Similar to influence spread, the running time grows with k. Although, some exceptions like IMM [90] and TIM [89] are also presented.

3. **Scalability: Running Time and Memory Consumption.** Scalability of an IM algorithm is measured in terms of both running time and memory consumption. Thanks to the highly efficient self-avoiding random walks generation on GPU(s), algorithms like [134, 135] run several orders of magnitude faster than its CPU's counterpart as well as the state-of-the-art methods. These methods take only seconds on networks with billions of edges and can work on even bigger networks by stretching the data across multiple GPUs.

4. **Robustness: Stability.** Robustness is equally important aspect of performance measure of an IM algorithm. An algorithm is called as robust, the optimal seed set does not change much when a slight

change occurs in diffusion model. The robustness evaluation in the literature is inadequate because it only focuses on the performance of IM algorithms on different social networks with various structures. In addition to this, it is also important to evaluate IM algorithms for various influence probability settings because influence probabilities are key components of influence networks, and they can directly affect the performance of IM algorithms. There are some works [43, 80, 136, 137] have been done to improve the robustness of the IM problem.

TABLE 2.6 concludes the performance of existing IM and context-aware IM algorithms with respect to above-discussed performance metrics and some observation are given as follow.

- None of the existing work can attain all four performance measure at the same time. Therefore, a careful selection of algorithm needed based on the required application.
- Most of the simulation-based algorithms have lower time efficiency and not scalable to large-scale networks. This is because these algorithms uses time-consuming MC simulations. All the simulation-based algorithms have good quality seed set. The memory consumption of these algorithms are not much high because only need to store sampled possible world and marginal gain of nodes.

- The time efficiency of scoring-based algorithms are much higher than simulation-based and much lower than sampling-based algorithms. This is because the score estimation process needs to scan the network several times. The seed quality have no theoretical guarantee on the approximation ratios of the score estimation algorithms, the quality of results is generally high in practice. These algorithms only need to store the information related propagation paths and score of nodes. Therefore, memory footprints are lowest among the existing algorithms. These methods have low memory footprint and can produce high quality results when good score estimation functions are used.

- Sampling-based algorithms have much higher time efficiency than simulation-based algorithms. The running time of the sampling-based algorithms are determined by the sample size. These methods are sensitive to influence probabilities and usually have high memory overheads.

- There are very few robust algorithms like IRIE, EASYIM are present in the literature. Both IRIE and EASYIM are insensitive to influence probabilities. This is because the influence score estimation process only performs arithmetic computations, and the computation time is independent of influence probability values.

TABLE 2.7: The Statistical Information of the Real-world Network Datasets

Dataset	$\|V\|$	$\|E\|$	$D_{avg}(N)$	Type
Dolphin [138]	62	159	5.12	Sparse
NetInfective [139]	410	17298	84.38	Dense
BHOSLIB [140]	450	83198	396.76	Dense
Email-Eu-core [141]	1005	25571	50.88	Dense
NetScience [142]	1589	2742	3.45	Sparse
Ca-GrQc [141]	5242	14496	5.53	Sparse
Gnutella08 [143]	6301	20777	6.59	Sparse
NetHEPT [24]	15233	31398	4.12	Sparse
Astro-ph [140]	16045	121250	15.11	Sparse
AS-22JULY06 [140]	22962	48435	6.03	Sparse
Gnutella30 [143]	36682	88328	4.81	Sparse
NetPHY [144]	37154	174161	9.38	Sparse
CM [38]	40420	175692	8.69	Sparse
NS [38]	1588	2742	3.45	Sparse
HT [38]	6360	15751	4.95	Sparse
Higgs Twitter [145]	456626	14855842	65.06	Dense
Football [146]	115	613	10.66	Sparse
Celegansneural [147]	297	2148	14.46	Sparse
USAir97 [148]	332	2126	12.80	Sparse
Political blogs [149]	1490	16718	22.44	Sparse
Amazon [150]	2880	3904	2.71	Sparse
Power [147]	4941	6594	2.66	Sparse

2.6 Datasets

Evaluation of influence maximization algorithms requires benchmark networks. There are widely used real-world networks compiled by network scientists specifically for social network analysis. Statistical information of network datasets use for experiments is given in Table 2.7.

Dolphin: Dolphin[1] [138] is an undirected social network of frequent associations between 62 dolphins in a community living off Doubtful

[1]http://www-personal.umich.edu/ mejn/netdata/dolphins.zip

Sound, New Zealand. The network was divided into groups depending on the association patterns of dolphins.

NetInfective: NetInfective[2] [139] network describes people's behavior during face to face interaction at INFECTIOUS exhibition.

BHOSLIB: BHOSLIB[3] [140] network is the collection of BHOSLIB.

Email-Eu-core: Email-Eu-core[4] [141] network was generated using email data from a large European research institution. They have anonymized information about all incoming and outgoing email between members of the research institution. There is an edge (u, v) in the network if person u sent person v at least one email. The e-mails only represent communication between institution members (the core), and the dataset does not contain incoming messages from or outgoing messages to the rest of the world.

NetScience: NetScience[2] [142] is a co-authorship network of scientists working on network theory and experiment, as compiled by M. Newman in May 2006. The network was compiled from the bibliographies of two review articles on networks, M. E. J. Newman, SIAM Review 45, 167-256 (2003) and S. Boccaletti et al., Physics Reports 424, 175-308 (2006), with a few additional references added by hand.

Ca-GrQc: Ca-GrQc[5] [141] is a collaboration network is from the e-print arXiv and covers scientific collaborations between authors papers

[2]https://arxiv.org/
[3]http://networkrepository.com/networks.php
[4]https://snap.stanford.edu/data/email-Eu-core.html
[5]http://snap.stanford.edu/data/ca-GrQc.html

submitted to General Relativity and Quantum Cosmology category. If an author i co-authored a paper with author j, the graph contains a undirected edge from i to j. If the paper is co-authored by h authors this generates a completely connected (sub)graph on h nodes.

Gnutella: Gnutella08[6] [143] and Gnutella30[7] [143] are peer to peer file sharing network snapshots. A sequence of snapshots of the Gnutella peer-to-peer file sharing network from August 2002. There are total of 9 snapshots of Gnutella network collected in August 2002. Nodes represent hosts in the Gnutella network topology and edges represent connections between the Gnutella hosts.

NetHEPT: NetHEPT[2] [24] is an academic collaboration network from the ”High Energy PhysicsTheory” section of arXiv form 1991 to 2003, where nodes represent the authors and each edge in the network represents one paper co-authored by two nodes. It contains 15233 nodes and 58891 undirected edges (including duplicated edges).

Astro-ph: Astro-ph[8] [140, 141] (Astro Physics) collaboration network is from the e-print arXiv and covers scientific collaborations between authors papers submitted to Astro Physics category. If an author i co-authored a paper with author j, the graph contains a undirected edge from i to j. If the paper is co-authored by h authors this generates a completely connected (sub)graph on h nodes.

[6] https://snap.stanford.edu/data/p2p-Gnutella08.html
[7] https://snap.stanford.edu/data/p2p-Gnutella30.html
[8] http://nrvis.com/download/data/misc/astro-ph.zip

AS-22JULY06: AS-22JULY06[9] [140] is the collections of miscellaneous networks.

NetPHY: NetPHY[2] [144] is constructed from the full paper list of the "Physics" section of the arXiv website.

Co-author: Co-author networks[10] [38] consists of three citation network in the field of Condensed Matter(CM), Network Science(NS), and High-Energy Theory(HT). To identify overlapping users in co-author dataset, we use authors name for matching. The number of overlapping users are 2860, 90, and 517 for network pairs CM-HT, HT-NS, and CM-NS respectively.

Higgs Twitter: Higgs Twitter dataset[11] [145] extracted from Twitter network based on user activities on elusive Higgs boson discovery between 1st and 7th July 2012. The Twitter network is follower-follow relationship network (FN) and it consists three types of interaction networks: Retweet network (RTN), Reply network (REN) and Mention network (MEN). The authors of [145] perform experiments on different periods and divide network based on these experiment time period. The experiment periods are: before 1 PM GMT on 2nd July (Period I), after 2nd July and before the announcement on 4th July (Period II), and after announcement to 7th July (Period III).

[9] http://networkrepository.com/as-22july06.php
[10] https://www.cise.ufl.edu/research/OptimaNetSci/tools/id_inter.html
[11] http://snap.stanford.edu/data/higgs-twitter.html

Football: Football[12] [146] is a football game network. The node set represents American Division IA colleges and edges denotes matches played between them.

Celegansneural: Celegansneural[13] [147] is a neural network of C. Elegans. Nodes represent neurons, and edges denote connections by either a synapse or a gap junction.

USAir97: USAir97[14] [148] is a US airline network where nodes and edges represent airports and the connectivity between airports.

Political blogs: Political blogs[12] [149] is a network of hyperlinks between weblogs on US politics, recorded in 2005 by Adamic and Glance.

Amazon: Amazon web graph[15] [150] is a network of web pages from amazon.com and its sister companies.

Power: Power[13] [147] is an undirected network of power grid located in western states of the United States.

2.7 Application of Influence Maximization

In this section, various applications that incorporate influence spread are studied and discussed how the influence maximization are utilized in those

[12]http://www-personal.umich.edu/ mejn/netdata/
[13]https://neurodata.io/project/connectomes/
[14]http://vlado.fmf.uni-lj.si/pub/networks/data/
[15]https://icon.colorado.edu/#!/networks

systems.

2.7.1 Influence Maximization in Events

Consider the following case: You are asked to host an event so as to promote either your company or a specific product. However, out of the 500 people in your network, your budget only allows you to invite 100 of them. Who will be on the guest list of your upcoming event, so as to maximize the visibility of your event?

Behind this rather simple and "school made" problem, some companies already rely on this question. For instance, video game developers have nowadays embraced the benefits of marketing through social networks such as YouTube. The context is the following: video game developers are facing a rise of the costs of developing games, mainly due to the fact that games have to be graphically enhanced through time; hence more and more lines of code are required to fully exploit the hardware capabilities. Therefore, the development teams have to be enlarged with valuable staff. The price of a single game staying relatively stable, the amount of sold games has to rise in order to follow the rise of costs and keep the same level of profitability. This is the reason why video games developers are investing more and more in marketing, and are exploring new marketing techniques.

Lately, one of this new techniques was to invite YouTubers (owner of a famous YouTube channel) to a special event prior to the release date of the game. Each YouTuber is then allowed to record some exclusive footage of the game he or she will be able to upload on his channel, generating views and ultimately some income via the YouTube monetization policy. These videos participate largely in the advertising of the video game and are really cheap compared to some other advertising vectors, such as TV commercials. If one could be able to extract the network generated by the different YouTubers, the usefulness of our "seeds selection" algorithms would become clearer. All the considered YouTubers represent the possible set of seeds, and the algorithm returns the chosen number of seeds so as to maximize influence through the network, ultimately returning the list of YouTubers to invite, maximizing the marketing coverage.

Even if this example is precise, one could imagine to extend this technique to other products and services. YouTubers gather millions of subscribers and could leverage an interesting income over cost ratio. Eventually, we believe that those techniques could become part of a possible mean of advertising among the offer of marketing companies. We can also extend the use of influence maximization regarding events. For instance, one could host an event for a specific product and the only way of participating to this event is through an application system linked to a Facebook account. At this point, we make sure that each applicant is sharing its useful information in order to build an efficient network. After

the application period is over, running the influence maximization algorithms trough the network composed of all applicants would return the guest list maximizing marketing coverage.

In a complete opposite way, a company could also base the decision to participate to a specific event based on its potential marketing coverage. Consider for example large forums, job fairs, etc. It could also be useful for a famous musician to know which music festivals to attend in order to maximize his coverage. Those examples are just a small glance of what could be a whole new marketing and coverage technique. What is important to remember is: for a specific occasion for which the number of applicants is higher than the total amount of guests and for which a specific network can be computed, IM techniques will attempt to select the best possible set of guests.

2.7.2 Influence Maximization in Recruitment

One possible application that is not related to marketing concerns recruitment and head-hunting, especially for high profiles and experts. With the spread of Internet through the world, experts can very easily get in contact with each others and publish results on specific platforms. For instance, *R-bloggers* is a notorious forum regarding the programming language *R*. Analyzing this forum and extracting a network out of it could lead to the possibility of using the techniques previously described in

order to detect the most influential members of the forum. If we take the hypothesis that one person's influence throughout a community of experts is directly linked to her skills, influence maximization techniques could point out the most competent people.

This idea can also be extended to any group of people forming a community, since there is a high chance that one can derive a network from it. We directly see that regardless of the field of expertise, influence maximization can under certain circumstances be used in a recruitment goal.

2.7.3 Influence Maximization in Social Media Population Screening

In light of recent events, we have seen that a coordinated campaign on social media might be used to influence large groups of people. These techniques have been used to influence peoples in Indian general election, 2019 .

In response to those, social medias have decided to delete accounts whose purpose is solely to conduct mass influence. However, it is nearly impossible to detect those fake accounts without an initial use of screening algorithms who will point out possible fake accounts. On the other hand, it is probably unwise to let the decision to delete an account or not to a computer. Therefore, human intervention is still needed at some point in the process so as to decide whether to delete an account or not.

Say, you have a list of X possible fake accounts pointed out by the previous algorithm, and a limited number of full-time equivalent (FTEs)to answer that demand. In order to limit the influence of the fake accounts, one might want to begin with the ones that are the most influential. This is where influence spread algorithms can be useful, since you can consider the accounts listed in X as potential seeds for your seed set. Then, by extracting and ordering the marginal influence of each seed, you can easily derive the ”potential fake accounts with the highest influence”.

Of course, as fake account screening techniques are unknown and unique for each company, the latest idea could be incorporated in some other way who would fit the actual technique. Therefore, we believe this idea is worth considering.

2.7.4 Others

If we broaden our definition of network, there are still some uncovered fields. Consider for example a problem posed by the Ministry of Jal Shakti, India: given a water distribution network and data on how contaminants are spreading through it, what is the best way to allocate sensors so as to detect all possible contaminations?

It has been proven that influence maximization techniques are more efficient than classical techniques used to address this kind of problems

[28]. Similarly, we can extend this result to road traffic. Roads being the edges and crossroads being the vertices, IM could enhance large scale road blocks and could help in finding a fugitive for instance. Furthermore, the reasoning could be extended for the dispatching of security agents in crowded areas, such as concert halls or department stores.

2.8 Business Implications of Influence Maximization

In recent years, the importance of micro-blog services as a marketing tool has multiplied, and its usage has spanned into diverse areas. For example, suppose that we are trying to market a product, an idea, innovation or behavior within a population of individuals, the commonly used advertising channel for such use case is peddling, putting signboards or playing promotion music. A promising way that could be used for such promotions is to place ads (advertisements) on popular micro-blog services (e.g., Twitter) through the viral marketing channel. Especially, if a company wants to promote its newly produced camera by placing ads, the company can convince several key people in the social network (influencer's) to adopt or try the new camera first, then the company may utilize the diffusion effect over the network to increase the effectiveness of its marketing campaign. If we assume that convincing each key person to spread the news on the new camera costs money, then a very significant problem could be described as: Given a social network, how can we

identify the key people through which we can spread the promotions for the new product most efficiently?

Recent empirical advances have used new observational techniques as well as randomized experiments to identify influence and susceptibility in networks. These advances provide new opportunities for specifying more accurate, contextual influence models when using influence maximization to identify optimal targets of public policy interventions or business advertising. Our results suggest that the growing body of research on influence maximization needs to incorporate results and insight from the empirical literature on influence identification to become more realistic and practically applicable. There are also some other business implications of IM problem like revenue maximization, and profit maximization from the perspective of advertiser and network service provider both. Apart from business advertising, there are some other potential applications of IM problem such as political campaign or elections, trend analysis and sales predictions, network monitoring, counter terrorism efforts, epidemiology, contagion management, etc.

To conclude, the phenomenon of viral marketing has made enterprises sit up and take notice. Businesses are now including viral marketing techniques into their customer relationship management efforts to communicate with their existing and prospective customers. As a businessperson, it depends on your strategic efforts in creating a viral

marketing strategy that will have a lasting positive influence on your business.

2.9 Summary

Influence maximization problem is studied in three steps: detecting interesting topics, modeling diffusion process, and identifying influential spreaders. First the study covers the information about product, innovation, news, etc. Second discussed the diffusion mechanism to propagate the information on the network leading to application. Lastly, the influential users are identified based on various methodological principles. Influence maximization problem has been studied several inter-disciplinary domains, yet the problem is not solved satisfactorily and leaves us with number of challenging issues.