Chapter 2

Literature survey

This chapter discusses state-of-the-art influence maximization in multilayer networks. Research in viral marketing takes advantage of social networks to promote a product. A product can spread in a network by finding a small set of influential users. These small set of influential users recommend the product to their followers, those followers recommend it to their followers, and so forth, creating a cascade of recommendations.

Identification of influential users is actively studied in [38, 57, 11, 29, 12, 82, 55, 8, 69]. Maximizing the influence through seed nodes was first proposed by Domingos et al. [28] and formulated as an optimization problem by Kempe et al. [38]. Kempe et al. [38] propose a greedy algorithm to find the influential users in a network using monte-carlo simulations. This greedy algorithm is an NP-hard problem[18, 66]. Since influence maximization problem is NP-hard, several heuristic based [13, 15, 12, 14], meta-heuristics based [10, 65], sampling based [20, 43, 21, 59], location based [80, 62], and several approximation algorithms [44, 22] have been proposed.

Arora et al. [3] have performed an experimental study on some influence maximization problems and proposed a benchmark framework. Figure 2.1 presents the benchmark framework, and it comprises four core components: empirical setup, seed node selection strategy, evaluation, and insights. Empirical setup includes the methods to compare with the state-of-the-art datasets, parameter configurations, and diffusion models. The seed selection process is algorithm-dependent, and all the other components of the benchmark remain the same. Evaluation provides the targeted diagnosis on the algorithms and results comparison. The evaluation process estimates the different performance measures such as influence spread, running time, memory footprints, etc. Insights discuss the key takeaway points from the benchmark and analyze the algorithm's results, effectiveness, scalability, efficiency, and robustness.

As we discussed above, generally, the seed selection process in IM can be classified into four categories: simulation-based approaches, heuristic-based approaches, sampling-based approaches, and community-based approaches [49]. Simulation-based approaches iteratively compute the marginal gain of each node; from that marginal gain, they select the seed nodes. Heuristic-based approaches use to find seed nodes based on ranking metrics and topological features like paths, distances, etc. Sampling-based approaches use the network's sample and nodes' reachability from one node to another. Community-based approaches find the communities in the network; seed node selection will be made from the communities. Related work for the above four approaches discusses below, and Figure 2.2 explains the different categories of IM problems.

2.1 Simulation based approaches

These models use the monte-carlo simulations to find the influential nodes as they can accurately estimate the influence spread. Pedro and Mat [28] have introduced the influence maximization problem in a graph; the core issue here is to spread the information from one node to another. The authors modeled the problem using Markov Random Fields and proposed heuristic solutions. Kempe et al. [38] studies influence maximization as an optimization problem and reported it as an NP-hard problem. Authors in [38] propose the greedy algorithm for influence maximization. Due to submodularity, the greedy algorithm produces near-optimal solutions with a theoretical guarantee. However, it suffers from scalability in large networks due to monte-carlo simulations. To improve the efficiency of the greedy algorithm, Leskovec et al. [44] propose the CELF algorithm, which utilizes the submodular property of the diffusion function to reduce the number of Monte Carlo simulations. CELF is 700 times faster than the greedy algorithm. Inspired by the idea of CELF, Goyal et al. [29] have introduced an algorithm CELF++ which is 55% faster than CELF. Although these improved algorithms increased efficiency, they still have poor scalability for more extensive networks. Cheng et al. [21] propose the StaticGreedy algorithm, which strictly guarantees the submodularity of the seed selection process. This algorithm reduces the computational expenses dramatically by two orders of magnitude without loss of accuracy. Ohsaka et al. [59] presented a snapshot-based sampling algorithm PRUNEDMC. To avoid re-computation, it prunes the breadth-first search. In addition, to reduce the cost of monte-carlo simulations, it updates the outputs of monte-carlo simulations iteratively.

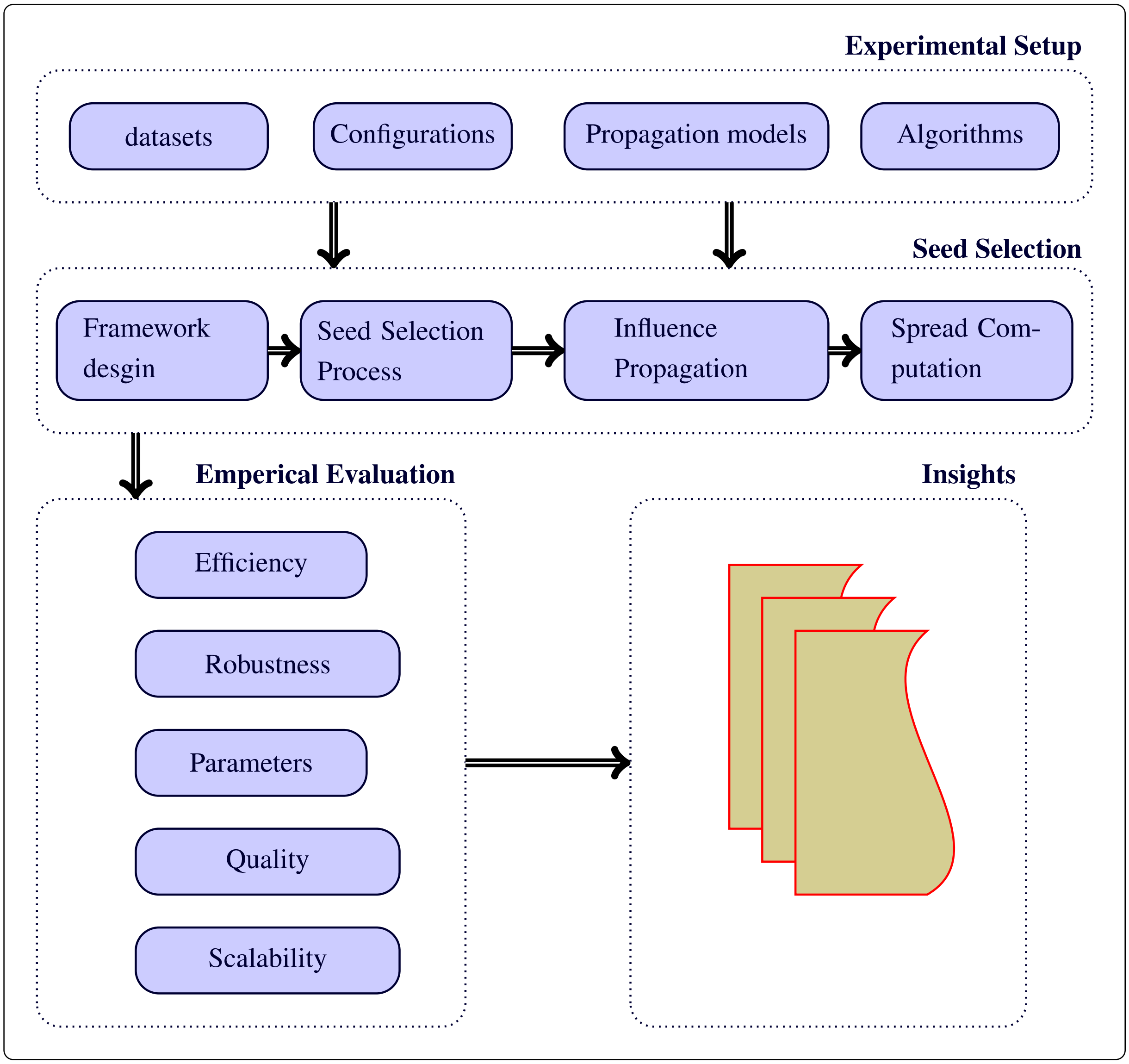

Figure 2.1: Benchmarking framework for IM [3]

2.2 Heuristic based approaches

To address the scalability issues, researchers propose various heuristic methods, including PMIA [12] and degree discount [11], to approximate the influence propagation using the node's local structures. Chen et al. [11] introduce an algorithm based on degree discount. The work is based on the observation that once a node selects as a seed node, that node is no longer available for seed selection. Therefore, a degree discount step is required for that node's direct neighbors. Finding seed nodes is difficult due to the non-availability of datasets, so the results will not be impressive. To overcome this, Kim et al. [39] proposed a multi-model deep learning model, influencer profiler (Infprofiler), which uses text and image information for finding seed nodes. Cai et al. [6] proposed Holistic Influence Maximization (HIM) query problem to find the minimum set of seed users who can cover all the targeted users in the network. Most IM problems focus only on online interactions and ignore offline interactions, but HIM focuses on both. However, these traditional algorithms ignore the community diversity of activated nodes in social networks and the time complexity of the IM.

2.3 Sampling based approaches

Nettasinghe et al. [58] have proposed a conditional influence estimation algorithm. It samples the seed nodes based on graph observation and transition probabilities. It uses the neighborhood size estimation algorithm with a variance reduction algorithm. Myriam [35] proposes MR-DSIN, a map-reduce based dynamic selection of influential nodes. It finds the seed nodes based on the graph sampling step to reduce the network's size. Zhang et al. [16] propose IMS influence maximization from these cascade samples. To achieve the optimization goal, IMS presents a solution to the network inference problem, i.e., learning diffusion parameters and network structure from the cascade data.

2.4 Community based approaches

Chen et al. [18] propose community-based influence maximization (CIM) to find the seed nodes from the communities and address the lack of community diversity and time cost in IM. It comprises three phases: community detection using hierarchical clustering; candidate generation, which aims to determine a set of candidate seeds based on community size and connectivity; and final seed selection from the communities. Kai Sheng et al. [64] propose a similar algorithm with label propagation for community detection called LPIMA. The Leader-

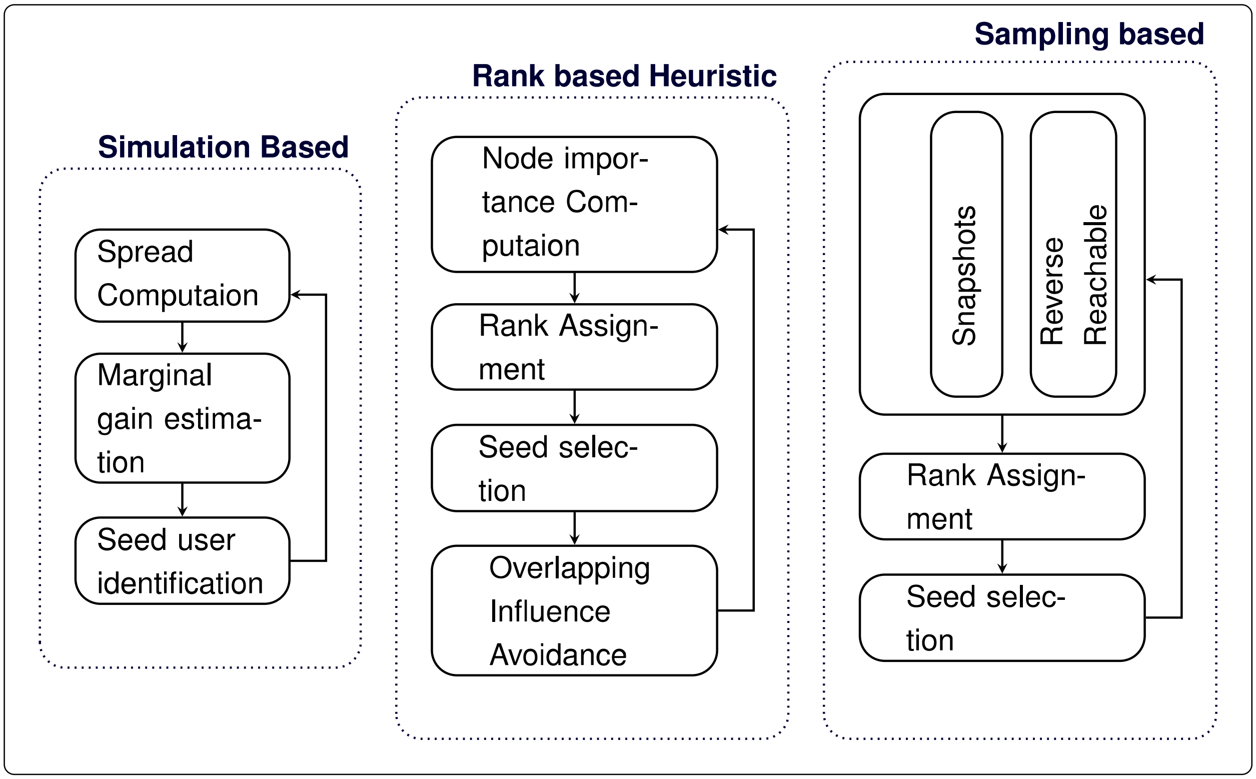

Figure 2.2: Seed selection process in various IM frameworks [67]

Rank algorithm uses to quantify community nodes and then assign candidate seed nodes based on quantified values; finally, the submodular characteristic uses to improve the greedy algorithm and select seed nodes. Jianxin et al. [46] propose Community-diversified Influence Maximization (CDIM) to find k seed nodes using the CPSP-Tree index measure, which measures the community-diversified influence and addresses a series of computational challenges. However, this method is based on theoretical models and is not practical for real-world scenarios.

However, the accuracy of selecting seed nodes is still a concern. Xiao Li et al. [48] propose a community-based seed selection algorithm (CSS) framework for location-aware influence maximization, which depicts finding the seeds by constructing PR-tree-based indexes. However, this framework is taking more network processing time after community partition. Wang Y et al. [77] propose the CGA algorithm (Community-based greedy algorithm), which reduces the processing time of the network. Noha et al. [2] propose the problem of temporal interaction-based community detection using clique structure. Xiaofei et al. [76] introduce the multi-community influence maximization (MCIM) problem to maximize influence by identifying seed users in multiple social communities of different properties and characteristics based on a total budget of seed users. Besides this, J. Guo et al. [30] propose influence maximization with community budget (IMCB). IMCB uses Newman Moore greedy modularity maximization to detect communities. Then a continuous greedy process and pipage rounding are used to find seed nodes from communities.

Table 2.1: Classification of various IM algorithms.

Category	Algorithm	Time complexity	Problem solving type	Base algorithm	Approximation	Diffusion Model
Simulation Based	Greedy	$O(kNMI)$	Spread simulation	–	$1-1/e-\varepsilon$	IC, LT
	Knapsack greedy	$O(N^5)$	Spread simulation	Greedy	$1-1/e-\varepsilon$	IC, LT
	CELF	$O(kNMI)$	Sub modularity	Greedy	$1-1/e-\varepsilon$	IC, LT
	CGA	$O(M+IM_C(N(Z-C)+K(C+N_C)))$	Community based	$1-e^{(-1/(1+\delta(C)))}$	–	IC
	LDAG	$O(Nt_\theta+Kn_\theta*m_\theta*(m_\theta+logn))$	Score estimation	–	N.A	LT
	BP-Greedy	–	Spread estimation	Greedy	$1-1/e$	IC, LT
	CELF++	$O(kNMI)$	Sub modularity	CELF	$1-1/e-\varepsilon$	IC, LT
Heuristic based	Simpath	$O(KlNP_\theta)$	Score estimation	N.A	LDAG	
	Degree discount	$O(KlogN+M)$	Degree based	High degree	N.A	IC, WC
	IM rank	$O(NTd_{max}logd_{max})$	Rank refinement	–	IC, LT	
	Cost degree	$O(M)$	Score estimation	–	N.A	IC, LT
	IPA	$O()$	Influence Path	PMIA	$N.A$	
	MIA/PMIA	$O(Nt_{i\theta}+Kn_{o\theta}*n_{i\theta}*(n_{i\theta}+logn))$	Influence path	SP1M	$1-1/e$	IC
	Diffusion degree	$O(N+M)$	Centrality based	High degree	N.A	IC
	EASYIM	$O(kD(N+M))$	Influence ranking	Greedy	N.A	IC
Sampling based	StaticGreedy	$O(\frac{kMN^2log_K^N}{\varepsilon^2})$	Snapshots	PMIA	$1-1/e-\varepsilon$	IC
	PRUNEDMC	$O(\frac{kMN^2log_K^N}{\varepsilon^2})$	Snapshots	Greedy	$1-1/e-\varepsilon$	IC
	TIM	$O(\frac{k(M+N)logN}{\varepsilon^2})$	Reverse reachability	$N.A$	$1-1/e-\varepsilon$	TM
Community based	CIM	$N.A$	Similarity based	$H_Clustering$	$N.A$	IC
	LPIMA	N.A	N.A	Community structure	N.A	IC
	CDIM	$O(n.m+n^2loglogn)$	Community detection	Shortest path distance	$1-1/e-\varepsilon$	IC, LT
	CSS	$O(\lvert C_h\rvert+k_{min}\lvert C_h\rvert\tau_\theta^c\tau_\theta^t)$	Community detection	Location based	$1-1/e$	IC
	CGA	$O(M+Z_MNT_P+MKT_p+K\lvert C_p\rvert T_p)$	Community detection	CIM	$N.A$	IC
	IMCB	N.A	Sampling based	Budget Allocation	$1-1/e$	IC
	MCIM	$O(n.K^2)$	Community detection	Budget Allocation	$1-1/e$	IC
	CINEMA	$O(M)$	Community detection	–	N.A	C^2,C^3
Multilayer networks	PCI	N.A	Centrality perspective	Degree	N.A	IC, LT
	CIM	$O(rNOMC+NOMClogNOMC)$	Clique structure	$N.A$	N.A	IC
	KSN	$O(nl(m+n)logn)$	Spread simulations	Greedy	$1-/e$	IC, LT
	MMNs	N.A	Spread simulation	Greedy	$1-1/e$	LT

Cao et al. in [7], where they first propose the community-based influence maximization algorithm OASNET (Optimal Allocation in Social Network). This algorithm assumes that the communities are independent of each other, i.e., influence can not transmit between two communities. Chen et al. [18] propose community-based influence maximization (CIM). Authors in [18] detect communities using a hierarchical clustering algorithm in the network. It selects the seed nodes based on the centroid of the communities. Kumar et al. [43] propose IM using Extended h-index and Label Propagation with Relationship matrix (IM-ELPR). IM-EPLR has four phases, first is the seeding phase to select the candidate seed nodes, and second is the label propagation phase to find the communities with the help of seed nodes. The third phase merges the communities using a relationship matrix, and the fourth phase selects the *k* influential nodes using the rank measure. Xuanhao et al. [17] propose community-based influence maximization in Location-based social networks (LBSN), which comprises three steps. The first step is to find the communities based on Spatio-temporal similarity measures between users. The second step finds the candidate nodes using the candidate selection algorithm. The third step uses two community detection algorithms to find final seed nodes. Jalayer et al. [34] have proposed a new greedy and hybrid approach based on a community detection algorithm for finding influential nodes. It is called Greedy TOPSIS and Community based algorithm abbreviated as GTaCB. The algorithm employs two well-known community detection algorithms. Chen et al. [11] use community structure to find the seed nodes for promoting the product. Kai et al.[64] propose the leader rank multi-label propagation algorithm (LRMLPA). LRMLPA finds the overlapping communities based on leader rank and selects the seed nodes from the communities. The above algorithms are computationally expensive due to the network processing time during community partition. Yuqing et al.[83] propose influential pricing nodes in the networks based on their expected influence spread. Authors in [83] design a function to characterize the relationship between price and expected influence.

2.5 Multilayer networks

After the advancement of the internet, users are not only restrained to a single layer. Due to the rapid growth of online social networks, users are connected with different layers, and each layer represents one of the many possible interactions. The four approaches mentioned above are traditional and mostly explored in single layer networks. However, these four approaches are also used in multilayer networks.

The influence maximization problem has yet to be fully explored in multilayer networks, but some efforts [4, 74, 73, 42] have done in this direction. Multilayer networks arise naturally in different applications such as social networks [41], biological networks [82], economy [55], transportation systems [8], temporal dynamics [69], and many others. Basaras et al. [4] use the power community index (PCI) to find the most influential nodes in multilayer networks. Degree heuristics considers only the degree of the node, but PCI intends to find the node with a dense neighborhood. Wang et al. [74] propose multilayer collective influence (MCI) to identify influential nodes, which utilizes topological and dynamic properties instead of the local degree of the node. Wang et al. [73] propose an algorithm, essential nodes determined based on CP tensor decomposition (EDCPTD), to find the most influential seed nodes. They applied CANDECOMP/PARAFAC(CP) tensor decomposition to get some significant factors, such as principle singular vectors. They produced quadruplets vectors to obtain hub and authority scores of each node across the layers. The seed selection in the greedy algorithm is not feasible for large networks due to the complexity of monte-carlo simulations. Therefore, Kuhnle et al. [42] propose a two-phase approximation algorithm called knapsack seeding of networks (KSN). The first phase runs parallelly in each layer and stores the list of activated nodes per seed node.

The second phase selects seed nodes based on the multiple-choice knapsack (MCK) problem. Venkata et al. [36] proposed a clique-based influence maximization (CIM) heuristic to find seed nodes in multilayer networks, and it consists of two steps. The first step is to find all the maximal cliques from the multilayer networks. The second step finds the seed nodes from the generated cliques.

Zhan et al.[81] propose a new model, a Multi-aligned Multi-relational network influence maximizer (M and M). It aims to extract multi-aligned and multi-relational networks (MMNs) using inter and intra-network social paths from aligned heterogeneous networks. For finding seed nodes, M and M extends the traditional linear threshold (LT) model to multi-aligned and multi-relational networks. A comparison of results for M and M is made in three different cases. In the first case is M and M algorithm selected seed nodes from Twitter and foursquare data sets. In the second case, iM and M selected seed nodes from only Twitter. M and M selected seed nodes from only foursquare in the third case. The first case performed better than the remaining cases. [1] propose a game theoretic model using Nash equilibrium to find the seed nodes on real datasets. It shows that none of the existing well-known strategies are stable, and at least one player is incentivized to deviate from the proposed system. On the graph topology, this method affects finding the influential nodes.

Kuhnle et al. [42] propose Knapsack seeding of networks (KSN) for influence maximization in multiplex networks. KSN works based on the influential seed finder (ISF) greedy algorithm [51]. Wang et al. [75] proposed influence maximization in multi-relational social networks (MRSN). Most IM problems focus on 1:1 relations, but MRSN is based on 1:N group relations. Qipeng et al. [50] formulated IM in multilayer networks as a multiobjective optimization problem and employed the classic non-dominated Sorting Genetic Algorithm II (NSGA-II) to find a set of Pareto-optimal solutions that provide a wide range of options for decision-makers. Due to the limited datasets, we can not extend single layer networks to multilayer networks. However, users have different influences on different social platforms, due to that, even if we extend single layer networks to multiayer networks, results may not be impressive.

The above literature study summarizes the work on influence maximization in single layer networks, and a few influence maximization approaches in multilayer networks. The algorithms mentioned above take more processing time and ignore the community structure in multilayer networks. Due to this, seed node selection may not be effective, and some parts of the network may never get the opportunity to receive the information when k (seed node-set) is limited. These shortcomings are addressed in our proposed algorithms CIM, SIM, CBIM, and K++-Shell decomposition.