Chapter 1

Introduction

A social network is a platform that provides a way to communicate, share new ideas, and connect with new friends. Online social networks have become powerful tools; they effectively connect people and make them active. In addition, social networks help in massive data dissemination and propagate ideas and innovation to influence a large number of people in a short time [41]. Of late, people spend a lot of time on the internet, and 60-70% of their time spend on social networks such as Facebook, Instagram, and Twitter [61]. This exponential growth in the usage of social networks is the genesis of viral marketing.

Nowadays, various web technologies and online social networking services have emerged. Social networks allow users to have connections (friends), and they facilitate sharing thoughts among friends through comments, pictures, etc. Around 68% of online users have a social profile to get news and connect with friends, family, or other interesting users. Many of these users form or join online communities, and they are more powerful marketing tools than regular traditional advertisements in the press and media.

With the advancement of mobile internet, awareness among users, and increasing usage of online tools, social networks have become critical places for companies to market their products. Online Social networks (OSN) attract billions of users to share information and activity. Popular social network sites are used for product marketing. In 2020, approximately 3.6 billion users were using social media, and this is projected to increase to 4.41 billion users by 2025. Of late, the expenditure on OSNs for advertisements surpassed traditional advertisements, which is over $50 billion [38]. This tremendous usage of social media has many opportunities for companies to promote their products through viral marketing. Social networks help track the

users' interests and behavior for product marketing and brand communication.

From the product marketing perspective, a user is not confined to a single social network. A user can have varying influential capacities on different social networks. So, influential social network users can help companies to promote their products. An influential user can act as a bridge between the company and the customer. When a company wants to promote a new product, it may have yet to gain prior knowledge about the market. Investing a significant amount to market the new product may not be possible. Even after investing a huge amount in marketing, the revenues may be poor for the product. Therefore marketing strategy should have qualitative decisions. Selecting the seed nodes (Influential users) is essential to promote its new product. For example, a startup may develop an application for online social networks and wants to promote it on the same platform. Assuming a lower budget to promote its application, a few influential social network users can be selected to gift the product with a request to review it. If influential users are selected appropriately, they can influence their friends and followers with their positive reviews. Selecting users who can maximize the information spread in social networks is called Influence Maximization.

Nevertheless, nowadays, individual users interact with their friends in a more complicated manner than ever. As we know, the single graph model has been used to represent social networks. After the advancement of the internet and its usage, most users are active on multiple social network platforms. For example, generally, people use Whatsapp or Hike for communication with family members, Facebook for communication with friends, Twitter for sharing news or current issues, LinkedIn for job search or professional posts, and Tiktok for entertainment videos [32]. It is easy to represent each scenario in the graph model separately.

However, single layer networks are incapable of dealing with multiple social accounts; therefore, multiple relationships between the same social users may not be representable effectively. Research interest in finding influential nodes in multilayer networks for viral marketing has been gaining popularity in recent years. Finding influential nodes in multilayer networks for various kinds of interaction is challenging. However, many issues should be addressed to fully utilize these social networks as marketing and information diffusion platforms. To address this, we present some influence maximization algorithms in multilayer networks in this dissertation.

The origins of multiple disciplines and associated tools to study the multilayer framework started decades ago; now, the study of multilayer networks has become one of the most important directions in network science. From the social network perspective, Kivela et al. [40] first proposed the multiple types of relationships, change in time, and other types of complications. Zaynab et al. [31] discussed the uses of multilayer networks in bio-medicine. [31] review the different aspects and terminologies of multilayered networks, and they discuss the variant applications of the multilayer framework.

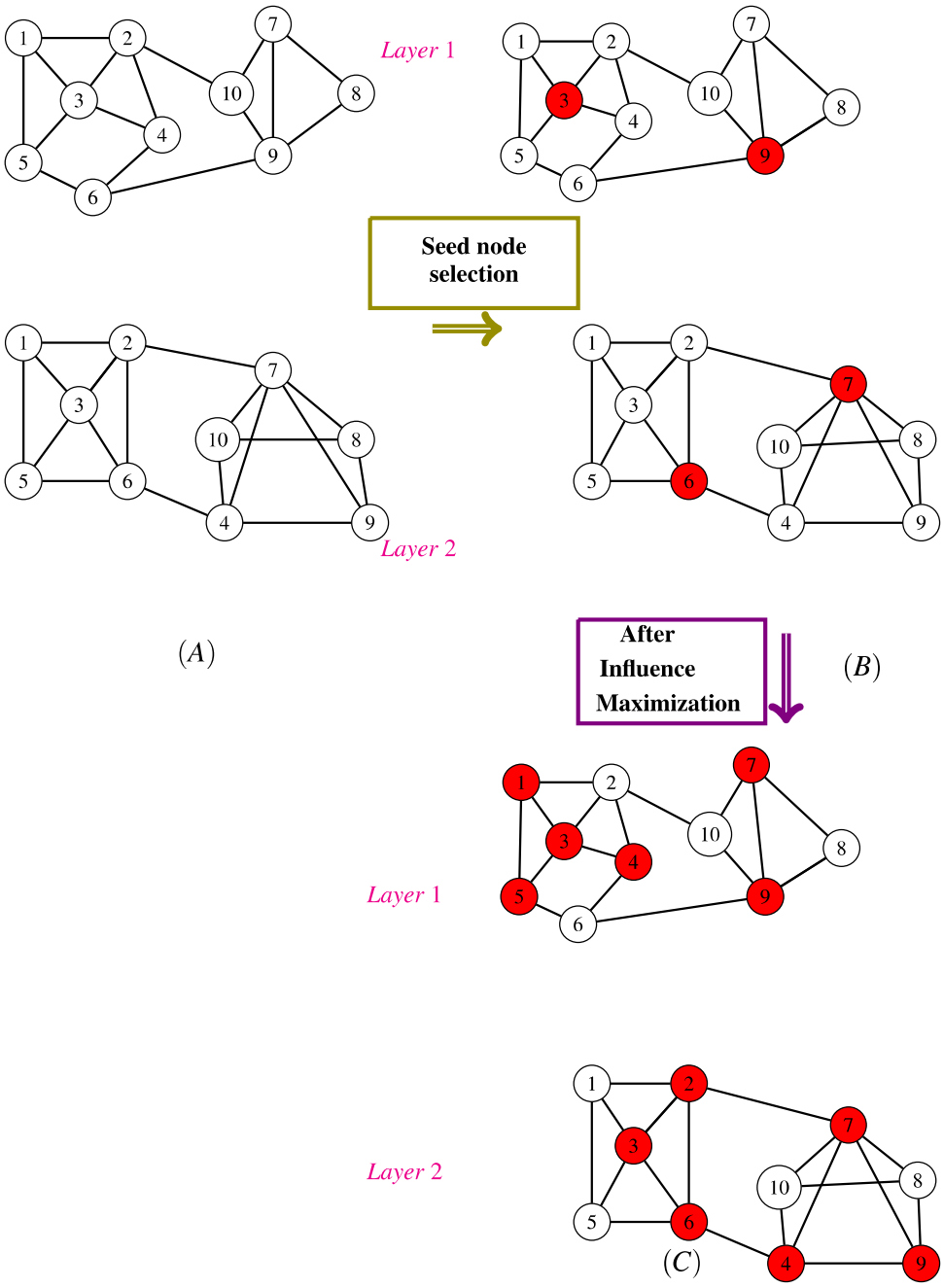

Figure 1.1: Overview of IM model

1.1 Influence Maximization

Influence maximization (IM) is the problem of finding a small set of influential nodes that can maximize the spread of influence. This section discusses influence maximization in multilayer networks. This dissertation uses two well-known propagation models for IM, i.e., the independent cascade (IC) model and the linear threshold (LT) model. Multilayer influence maximization refers to finding the most influential nodes from all the layers, called seed nodes. The expected influence spread according to any propagation model is maximized in multilayer networks by activating them. The propagation model plays a crucial role in the analysis of influence maximization. In both these models, a node is assumed to be either in an active state or an inactive state [38].

To achieve the goal of viral marketing, it is essential to find influential nodes for maximizing the information spread. Generally, systems are often represented as networks in various areas like sociology, physics, biology, computer science, etc. In these areas, influence maximization is vital for spreading information.

Figure 1.1 presents the overall flow of influence maximization in a multilayer network. A multilayer network has two layers, and each layer has ten nodes. In layer 1, nodes 3 and 9 are selected as seed nodes. In layer 2, nodes 6 and 7 are designated as seed nodes. After running any one propagation model, nodes 1, 2, 4, 5, 7, and 9 are activated in layer 1, and nodes 2, 3, 4, 6, 7, and 9 are activated in layer 2.

1.1.1 Independent Cascade Model

In the independent cascade model, when a node u^l becomes active at time step t, it will get one chance to activate its inactive neighbor $(v^l, w^l, x^l ...)$ at time step $t+1$ with propagation probability $p_{(u^l v^l),(u^l w^l),(u^l x^l)...}$. In the Independent cascade (IC) model, propagation probability lies between 0 and 1, also called success probability. Regardless of its success in activation, the same node will never get another chance to activate its inactive neighbors.

1.1.2 Linear Threshold Model

In the linear threshold model, every node is associated with an activation threshold between 0 and 1. At any time step t, if the sum of the incoming edge weights is greater than the activation threshold of the node, then the node enters into an activated state. The sum of any node's all incoming edge weights is assumed to be at most 1.

A node v^l is influenced by each neighbor w^l according to b_{w^l,v^l} such that

$$\sum_{w^l \in nbrs(v^l)} b_{w^l,v^l} \leq 1$$

A node v^l becomes active when at least (weighted) θ_{v^l} fraction of its neighbors are active, i.e.,

$$\sum_{w^l \in actnbrs(v^l)} b_{w^l,v^l} \geq \theta_{v^l}$$

In both models, influence propagation continues until no new user turns active.

1.1.3 Submodularity

As both IC and LT models satisfy the natural diminishing returns property, they are submodular. The influence of a node decreases when it is part of a more extensive seed set because other seed nodes can partially take over its role in propagation in the network.

Let χ be a collection of coin flips on edges and $R(v^l,\chi)$ be the set of all nodes which can be reached from node v^l on a path consisting of totally live edges; we can estimate approximate influence spread $\sigma(S)$ of seed set S [52].

$$\sigma(S) = |\cup_{v^l \in S} R(v^l,\chi)|$$

where $\sigma(S)$ is the expected number of nodes influenced by seed set S after applying one of the propagation models.

Definition 1 (Monotone)**.** f is monotone, if $\sigma(A) \leq \sigma(B)$ whenever $A \subseteq B$. i.e. $\sigma(S \cup \{v^l\}) \geq \sigma(S)$. Influence spread is more in the super set when compared to the subset,

Definition 2 (Submodularity)**.** The function $\sigma : 2^V \rightarrow \Re$ is sub modular if for all $X,Y \subset V$

$$\sigma(X) + \sigma(Y) \geq \sigma(X \cup Y) + \sigma(X \cap Y) \tag{1.1}$$

If all the elements are submodular functions, then we say that a collection σ of functions from $2^V to\ \Re$ is submodular [54].

Monotonicity is where the effect of a more extensive set L on node u is more potent than that of a smaller set K. Submodularity is the condition where the product (diminishing returns) of adding a node u to a smaller collection K is more than the effect of adding the same node u to a more extensive set L, $(K \subset L)$ [57].

Lemma 1. A function $\sigma : 2^V -> \Re$ is sub modular Iff

$$\sigma(A \cup \{v^l\}) - \sigma(A) \geq \sigma(B \cup \{v^l\}) - \sigma(B) \tag{1.2}$$

for all $A \subset B \subset B \cup \{v^l\} \subset V$, and $v^l \in V \setminus B$.

Proof. From definition 2, assume $X = A \cup \{v^l\}$, $Y = B$ and substitute in Eq. 1.1

$$\sigma(A \cup \{v^l\}) + \sigma(B) \geq \sigma(A \cup \{v^l\} \cup B) + \sigma(A \cup \{v^l\} \cap B)$$

$$\sigma(A \cup \{v^l\}) + \sigma(B) \geq \sigma(B \cup \{v^l\}) + \sigma(A)$$
$$\sigma(A \cup \{v^l\}) - \sigma(A) \geq \sigma(B \cup \{v^l\}) - \sigma(B)$$

Hence, the proof that Equation 1.2 satisfies the submodularity property. For the IC and LT models, σ is monotone, and submodular [47].

□

1.2 Motivation

Identification of influential nodes in a network has many practical applications. One of the primary motivations for influence maximization is viral marketing. For example, viral marketing is a good strategy to promote a product. The viral marketing strategy will identify the most influential nodes, convince them to adopt, then endorse that product and make them influence their followers in the network for viral spreading.

Social media users do not confine to a single social network; they participate in different social networks. In such a case, a user can propagate the information across different social networks, i.e., multilayer networks [41]. However, the majority of the existing studies focus on influence maximization in single layer networks. Single layer networks ignore critical factors such as user engagements across the networks, the network of networks, etc. After the internet evolution, users are excited to have multiple social accounts, such as Facebook, Twitter, etc. Therefore, we need to consider multilayer networks, as many users are actively engaged in simultaneously propagating information, ideas, and innovations across the networks. Moreover, users have different influences on different social platforms. For example, a user may have more

followers on Twitter than on Facebook, and vice versa. Therefore we need to consider multilayer networks as many users are actively engaged in sharing information across the networks simultaneously. Despite this, relatively little research is dedicated to IM in multilayer networks. At this point, detecting influential nodes precisely in multilayer networks is vital for spreading information. This concept is a challenging yet unexplored task. In this thesis, we design and develop a few algorithms for IM problems in multilayer networks.

1.3 Thesis Contributions

The main contributions of the thesis are present in this section. The thesis aims at finding the most influential nodes in multilayer networks. The major contributions are as follows:

- CIM: Clique-based heuristic for finding influential nodes in multilayer networks.
- SIM: Similarity-based community influence maximization in multilayer networks.
- CBIM: Community-based influence maximization in multilayer networks.
- K++-Shell: Influence maximization in multilayer networks using community detection.

1.4 Thesis outline

✓ Chapter-1: Discusses an overview of the influence maximization in multilayer networks, influence maximization models, motivation of the thesis, thesis contributions, and the thesis outline.

✓ Chapter-2: Discusses the literature study of the influence maximization in multilayer networks. This chapter briefly discusses the various types of identifying the most influential nodes in social networks. This chapter discusses the benchmark models for Influence maximization. It discusses the state-of-the-art simulation-based approaches, heuristic-based approaches, sampling-based approaches, and community-based approaches for finding influential nodes. It also discusses some models to identify influential nodes in multilayer networks.

✓ Chapter-3: This chapter discusses a novel algorithm, CIM, a clique-based heuristic for finding influential nodes in multilayer networks. CIM finds the seed nodes using a clique-based structure. It finds all the maximal cliques in the multilayer networks. After that, seed nodes select from these generated cliques. For finding the seed nodes, CIM proposes

the seed selection process; the selection process consists of four cases, and the seed nodes will be identified based on these cases. In addition, we also propose ignoring *noted nodes* to save the time complexity for the influence maximization and remove the information redundancy. CIM ignores the nodes for information spreading if they are already seed nodes in any of the one layers.

✓ Chapter 4: One of the disadvantages of CIM is its insensitivity to the community structure. This thesis proposes SIM, Similarity-based community influence maximization. SIM overcomes the disadvantages of CIM. SIM is an incremental community-based algorithm for influence maximization. Generally, individuals in a community interact frequently and are more likely to influence each other. SIM algorithm uses the community structure to find the most influential nodes in the multilayer networks. SIM comprises four stages. In the first stage, we propose the weighting index (WI) metric and calculate WI for each node. In the second stage, SIM uses WI and Jaccard similarity to find the initial community set. At this stage, a large number of small communities will be generated. It violates the fundamental characteristic of a community. In the third stage, SIM consolidates some small communities using the cut similarity metric and forms the final community set. In the fourth stage, the seed node selection process will be done from each community using a quota-based approach based on the size of the community.

✓ Chapter 5: One of the disadvantages of SIM is the lack of quality communities and seed nodes. This thesis Elucidates CBIM, community-based influence maximization in multilayer networks. CBIM is another community-based influence maximization algorithm to find much influential nodes for spreading information. CBIM overcomes the disadvantages of SIM. CBIM also finds the most influential nodes, and it also uses community structure to find seed nodes. The CBIM algorithm has two stages: The first is for community detection, and the second is for seed node selection. In the first stage, CBIM finds the initial communities. Algorithms use the degree of a node and the dice similarity index for finding initial communities. Then, CBIM merges some initial communities to form the final community set. We propose merging index (MI) as the criterion to merge some of the small communities. MI calculates based on a community scale and community conductance. This thesis proposes an edge weight sum (EWS) in the second stage, similar to the weighting index (WI). CBIM computes EWS for each node in all the generated communities and ranks the nodes based on EWS. CBIM uses a quota-based approach to select the seed node set from the communities based on the nodes' EWS ranks.

✓ Chapter 6: SIM and CBIM take more processing time for communities detection and seed node selection. This thesis proposes a K++-Shell decomposition algorithm. It reduces the time complexity and improves the performance in influence spreading than previous algorithms. The K++-Shell decomposition algorithm aims to find the most influential nodes in multilayer networks. K++-Shell decomposition is an improved version of the K-

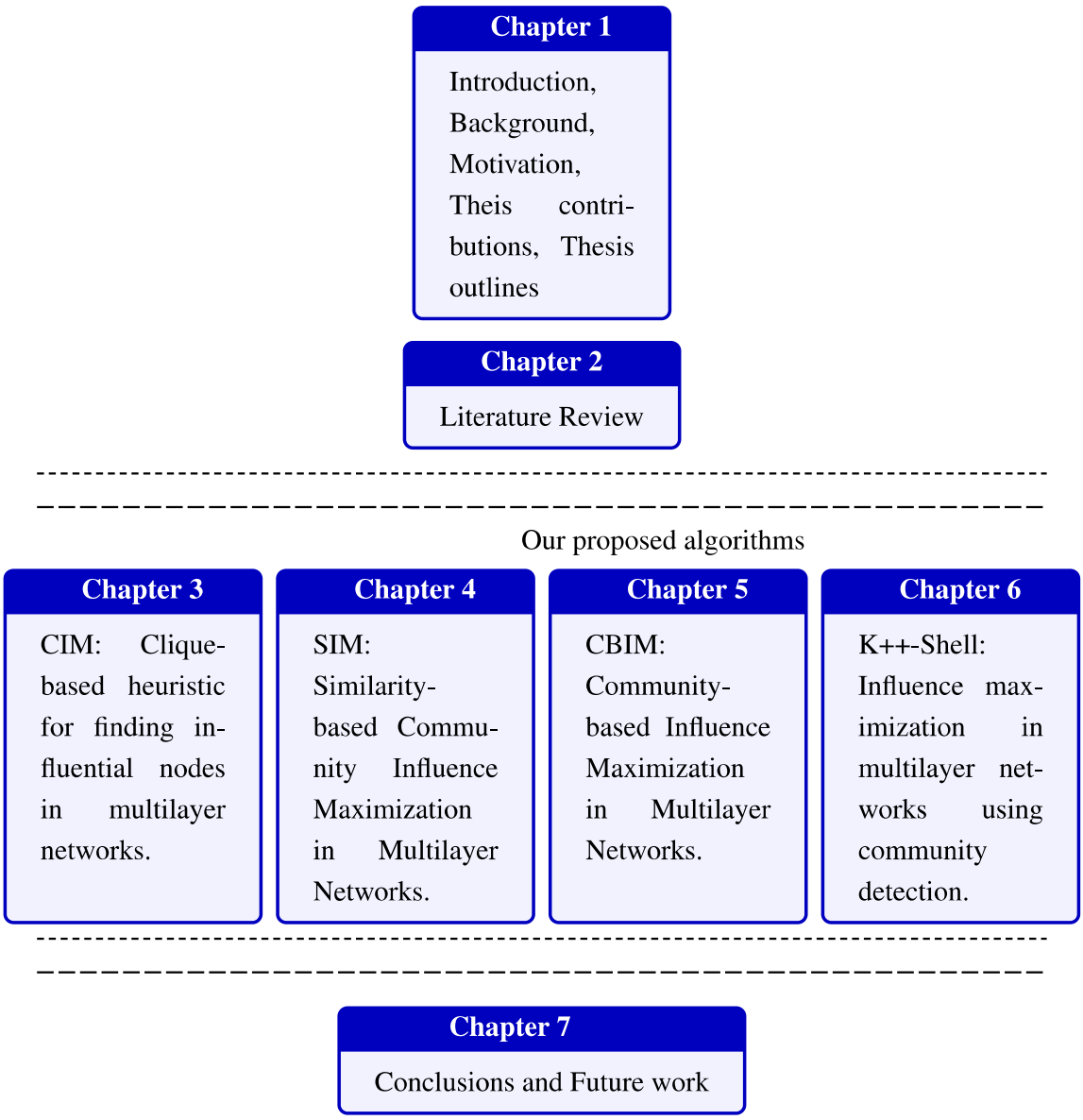

Figure 1.2: Thesis organization

Shell decomposition algorithm. Generally, the K-Shell decomposition algorithm prunes the nodes based on degree and places the pruned nodes in the appropriate buckets. But it ignores the critical aspect, i.e., all the nodes in the highest bucket may not be influential, and some nodes in the lower bucket may be much more influential as they may have rich neighbors. To address this problem, the K++-Shell decomposition algorithm rewards one point to each neighbor of a pruned node. Before K++ Shell decomposition, we find the communities using the label propagation algorithm in a multilayer network.

✓ Finally, Chapter 7 concludes thesis with few future directions