CHAPTER 3

GAME THEORETIC APPROACH

3.1 OVERVIEW

Social networks play a vital role in spreading influence by *word-of-mouth* behavior. Email networks, scientific collaboration networks and online social networking sites such as Facebook, LinkedIn, etc. are some examples for real world social networks. In a social environment, decision taken by an individual on a given issue is influenced by the behavior of their neighbors. Among all those neighbors, only a few have the ability to exert strong influence on others. They are called as *influential individuals*. The primary issue in this context is to find this subset of influential individuals of desired cardinality K, who can eventually influence maximum number of people in the network. This problem is referred to as top-K nodes problem that has got many important applications such as marketing a product or a technology, finding prolific authors in a research domain and so on.

Finding such influential nodes is an NP-hard problem computationally. Approximate algorithms to solve this problem widely use traditional social network analysis metrics such as degree centrality to select the initial set of influential individuals. These approaches treat each individual as an independent entity. In reality, individuals exist as communities and perform better in groups than in isolation. Game theory provides rich mathematical model to measure coalition strength among individuals. However, the state-of-art game theoretic algorithms are computationally expensive and do not provide scalable solutions for social networks of larger size.
The work expounded in this chapter is an attempt to describe a comprehensive game theoretic framework to find influential nodes, given a well-defined perspective in the social network. The proposed work in this chapter provides a tri-stage game theoretic solution mechanism with reduced time complexity for finding such influential nodes. (1) Initially communities are identified based on the influence spreading ability of the individuals, (2) Bridge nodes, which are proved to be effective in maximizing influence spread, are identified in each community, and (3) Influential nodes are selected among the bridge nodes.

### 3.2 INFORMATION DIFFUSION MODELS

Information diffusion models for spreading an idea or innovation through a social network, represented by a graph, considers each individual node as being either active (an adopter of the product or technology) or inactive. The nodes are assumed to switch from inactive to active state, but not vice versa. In this setting, each node’s tendency to become active increases monotonically as more of its neighbors becomes active. An initially inactive node \( v \) may become active as more and more neighbors of \( v \) become active after certain time period. The decision of \( v \) may in turn trigger further decisions by nodes to which \( v \) is connected. Two information diffusion models widely studied in the literature are (1) Linear threshold model and (2) Independent cascade model. This section briefly discusses about these models. Linear threshold model is adopted in the proposed approach to study the influence spread.

#### 3.2.1 Linear Threshold Model

Linear threshold model proposed by Granovetter (1978) and generalized by Watts (2002), initially assumes every node to be inactive. That is, no node has adopted the product or the technology. The flexibility of the
node $i$ to adopt a product (or technology) is captured by its threshold $\theta_i$. For each node $i$, let $N_i$ be its neighbors and $active(i) \subseteq N_i$ represent its active neighbors, who have adopted the product (or technology). Node $i$ is influenced by a neighbor node $j$ according to a weight $w_{ij}$ that are normalized such as $\sum_{j \in N_i} w_{ij} \leq 1$. The decision of a node $i$ to become active is based on a threshold function $f_i$ of the set of active neighbors of $i$ and a threshold $\theta_i$ chosen uniformly at random by node $i$ from the interval $[0, 1]$. The function $f_i : 2^{N_i} \rightarrow [0, 1]$ is defined as $f_i(T) = \sum_{j \in T} w_{ij}, \forall T \subseteq N_i$. The threshold $\theta_i$ represents the weighted fraction of the neighbors of node $i$ that must become active in order for node $i$ to become active.

Given a choice of thresholds and initial set of active nodes the information propagates at given time instance as follows: All the nodes that were active at time step $t-1$ remains active. At time $t$, an inactive node $i$ would be activated if the total sum of the weights of its active neighbors exceeds its threshold, that is $f_i(\text{active}(i)) = \sum_{j \in \text{active}(i)} w_{ij} \geq \theta_i$. The nodes thus activated are assumed to remain in the same state thereafter. This process stops when there is no new active node in a particular time interval.

3.2.2 Independent Cascade Model

This model is originally proposed by Lopez-Pintado (2008) has an important parameter called diffusion speed $\mu$. An active node $i$ in step $t$ activate its inactive neighbor $j$ to become active, at a probability rate $\mu$. In viral marketing, diffusion speed models the tendency of individuals to accept a product. Thus, the information spread is affected by diffusion speed, node degree, and the number of initial active nodes. If node $i$ succeed, then node $j$ will become active in step $t+1$; but whether or not $i$ succeed, it cannot make any further attempts to activate $j$ in the subsequent rounds. Again, the process runs until no more activation is possible.
3.3 GAME THEORETIC FRAMEWORK

Game theory provides a powerful mathematical tool that analyzes the strategic interactions among a group of individuals, which leads to better understanding of users’ complex behavior dynamics. It has been developed for understanding cooperation and conflict between individuals in many fields such as economics, politics, business, social sciences and biology. The games studied in the game theory are well-defined mathematical objects. A game consists of a set of players, a set of moves (or strategies) available to those players, and a specification of payoffs for each combination of strategies. In essence, each player has to decide a set of moves which are in accordance with the rules of the game that maximizes the rewards. The study on game theory is classified into two branches: (1) Non co-operative game theory, in which players work independently without assuming anything about what other players are doing. (2) Co-operative game theory where players co-operate with one another and work in unison to attain a common goal.

This study attempts to describe a co-operative game theoretic framework to find influential nodes in a social network. In the proposed framework, individuals in a social network are assumed to be players in a game. A co-operative game is defined as a pair \((N, \vartheta)\), where \(N\) represents set \(\{1, 2, 3, \ldots, n\}\) of \(n\) players and \(\vartheta\) represents the characteristic function that measures the coalition strength of players. Given a subset of players, \(S \subseteq N\), \(\vartheta(S)\) is a real number that gives the coalition strength of \(S\) players without help from other players \(N \setminus S\). \(\vartheta(N)\) measure the total coalition strength of all the \(N\) players. The following section describes Shapley value, the solution concept of co-operative game theory.

3.3.1 Shapley Value

An important requirement in a co-operative game is to divide the payoff fairly among the players, as the contribution of some players to the
game will be more than others. The concept of Shapley value, which was developed axiomatically by Shapley (1953), takes into account the relative importance of each player to the game in deciding the payoff to be allocated to the players. The Shapley value of the co-operative game \((N, \vartheta)\) is defined in Equation (3.1).

\[
\varphi(N, \vartheta) = \varphi_1(N, \vartheta), \varphi_2(N, \vartheta), ..., \varphi_n(N, \vartheta)
\]  

(3.1)

where \(\varphi_i(N, \vartheta)\) denotes the Shapley value for player \(i\). Given a subset of players \(S \subseteq N\) such that \(i \notin S\), the marginal gain of player \(i\), to the subset \(S\) is defined as \(\vartheta(S \cup \{i\}) - \vartheta(S), \forall S \subseteq N \setminus \{i\}\). The exact Shapley value of player \(i\), is defined in Equation (3.2).

\[
\varphi_i(N, \vartheta) = \frac{1}{n!} \sum_{T \in \Theta} [\vartheta(S_i(T) \cup \{i\}) - \vartheta(S_i(T))]
\]  

(3.2)

where \(\Theta\) represents set of all \(n!\) permutations of nodes in \(N\), \(T\) represents a single permutation in \(\Theta\) and \(S_i(T)\) represents the set of players that appear before node \(i\) in permutation \(T\). The concept of Shapley value guarantees fair allocation of the total payoff among the players based on four axioms: (1) Efficiency (2) Symmetry (3) Dummy and (4) Additivity.

**Efficiency:** This axiom requires that players precisely distribute among themselves the resources available to the grand coalition, as defined in Equation (3.3).

\[
\sum_{i \in N} \varphi_i(\vartheta) = \vartheta(N)
\]  

(3.3)

**Symmetry:** This axiom requires the following notion of symmetry: Players \(i, j \in N\) are said to be symmetric with respect to a game if they make the same marginal contribution to any coalition, that is for each \(S \subseteq N\) with \(i, j \in S\)
The symmetry axiom requires symmetric players to be paid equal shares. If players $i$ and $j$ are symmetric then $\varphi_i(\emptyset) = \varphi_j(\emptyset)$.

**Dummy**: The third axiom requires that zero payoffs be assigned to players whose marginal contribution is null with respect to every coalition. If $i$ is a dummy player, that is $\vartheta(S \cup \{i\}) - \vartheta(S) = 0$ for every $S \subset N$, then $\varphi_i(\emptyset) = 0$.

**Additivity**: Finally, it is required that the Shapley value be an additive operator on the space of all games, that is $\varphi(\emptyset + \omega) = \varphi(\emptyset) + \varphi(\omega)$.

### 3.4 PROPOSED STRATEGY

The objective of the proposed approach is to maximize the influence spread with reduced computational complexity by taking advantage of the small-world properties inherent in the social network. The small-world properties analyzed in this study are high clustering coefficient that indicates the presence of communities and weak ties (bridge nodes) that connect the communities. The classic study conducted by Granovetter (1973) has proved that weak ties are effective in maximizing the influence spread.

#### 3.4.1 Study of Nodes Influence

The social network is assumed to be weighted undirected graphs with edge weights represent the influence degree of a node on its neighbors. Initially, all the nodes in the network are assigned with random threshold $\theta \in [0, 1]$, and the edges are assigned with random weights $w_{ij} \in [0, 1]$. All the nodes are assigned with unique labels from the set $\{1, 2, \ldots, n\}$ and associated with an influence vector $I$ of size equal to the number of neighboring nodes. Each element of the influence vector of node $i$ represent its neighbor node $j$ that has value set to true if $i$ is influenced by node $j$, false otherwise.
Initially node with label $j$ is activated, and its influence on the neighboring nodes based on the linear threshold model is studied. This process is repeated for all the nodes in the network. The procedure for finding influential behavior of nodes is outlined in Algorithm 3.1.

**Algorithm 3.1 Node Influence Algorithm**

Input : Weighted Social Network $G(V,E,W)$, influence vector $I$ of nodes  
Output: Initialized influence vector of nodes  

1: $I_1 = I_2 = I_3 = \ldots, I_n = \emptyset$  
2: for each $v \in V$ do  
3:     for each neighbor $u_j$ of $v$ do  
4:         if ($u_j$ influences $v$) then  
5:             $I_v(u_j) = true$  
6:         else  
7:             $I_v(u_j) = false$  
8:     end if  
9:     end for  
10: end for  

### 3.4.2 Community Detection

The label propagation algorithm presented by Raghavan et al (2007) and adopted by Wang et al (2010) to detect communities based on influence degree of nodes is used in the proposed approach to detect communities. In Wang et al (2010) communities are formed based on the independent cascade model and a dynamic programming algorithm is used to find influential nodes. In this study, communities are initially formed based on the linear threshold model and influential nodes are identified based on the game theoretic approach.
The objective of the label propagation algorithm is to reassign the label of a node to another label that is assigned to maximum number of its active neighbors. This step is repeated until a steady state is reached, that is the number of reassignment falls below a threshold $\beta$. Finally, nodes with the same label are grouped together to form communities. Formally, the label of a node $v$ at iteration $t$ denoted by $l.v^t$ is reassigned as given in Equation (3.4).

$$l.v^t = \text{maxlabel}(l.u_1^{t-1}, l.u_2^{t-1}, ..., l.u_k^{t-1})$$

(3.4)

where $u_j$, is an active neighbor of node $v$ and the label of the node $u_j$ at time $t-1$ is represented by $l.u_j^{t-1}$. The set of active neighbors of node $v$ would be straightforwardly obtained from the vector, $l_v$. The procedure for detecting communities based on the nodes influence is given in Algorithm 3.2. The communities thus formed primarily consist of individuals capable of influencing each other.

---

**Algorithm 3.2 Community Detection Algorithm**

Input: Set of nodes assigned with unique community labels and threshold $\beta$

Output: M-communities consisting of influential nodes

1. while (number of label reassignments > $\beta$) do
2.     for each $v \in V$ do
3.         $l.v^t = \text{maxlabel}(l.u_1^{t-1}, l.u_2^{t-1}, ..., l.u_k^{t-1})$
4.     end for
5.     $t=t+1$
6. end while
7. Combine nodes with same labels into communities.
8. Let M be the number of communities formed

---

### 3.4.3 Role of Bridge Nodes in Influence Maximization

To exemplify the advantage of weak ties, consider Figure 3.1, that has two communities \{A-G\} and \{H-K\} of influential nodes. Choosing node
B as influential node would be more appropriate as it connects otherwise disconnected communities and can better maximize the influence spread than the other nodes in the network.

![Figure 3.1 Role of Bridge Node B as Influential Node](image)

With this motivation, the next step is to identify bridge nodes across the communities. The proposed Algorithm 3.3 returns bridge nodes across communities, given the set of communities identified in the previous step as input. This algorithm differs from the existing algorithms in the literature that explores only the link structure (cut vertices) to find bridge nodes. The proposed algorithm identifies bridge nodes within each community based on the number of communities it is connected. To avoid clustering of bridge nodes within a community, each community is associated with a bridge vector $BV$. The size of $BV$ vector for each community $C_i$ is set according to the number of nodes present in it.

Each element in the vector that represents a node $v$ within $C_i$ holds the number of communities the node $v$ is connected. Finally, a node in $C_i$ that connects to maximum number of other communities would be chosen as its bridge node, with ties broken arbitrarily.
Algorithm 3.3 Identifying Bridge Nodes across M Communities

Input: Communities of Influential nodes  
Output: Bridge nodes for each community  

1. Let M be the number of communities identified  
2. for (each community $C_i$ in G)  
3. \hspace{1em} $BV_{C_i}$=0  
4. end for  
5. for (i=1 to M)  
6. \hspace{1em} for (each vertex v in $C_i$)  
7. \hspace{2em} for (each neighbor $u_j$ of v, $j \in [1 - \text{deg}(v)])$  
8. \hspace{3em} if ($h_{u_j}[v] = 1$) and ($C.v \neq C.u_j$) // C.v- community of node v  
9. \hspace{3em} $BV_{C_i}[v] = BV_{C_i}[v] + 1$  
10. end if  
11. end for  
12. end for  
13. bridge[$C_i$]= argmax{$BV_{C_i}[v]$} \forall v in $C_i$  
14. end for  

3.4.4 Redefined Shapley Value to Identify Top-K Nodes

Let $b$ be the number of bridge nodes identified by the previous step and $B = \{1, 2, \ldots, b\}$ represents the set of bridge nodes. Node $i$ among the bridge nodes is considered to be more influential than node $j$, if $i$ connects to more communities than $j$. The influential nodes among the bridge nodes are then identified based on their Shapley value. As naïve Shapley value computation is intractable, the proposed approach thrives to reduce the computational complexity in two ways:

1) Top-K influential nodes are identified only among the bridge nodes $b$, where $b \ll n$ and $n$ is the total number of nodes in the network.

2) Traditionally, the characteristic function $\mathcal{S}(S)$ represents the expected number of nodes activated by the set of nodes $S \subseteq N$. The proposed approach redefines $\mathcal{S}(S)$ to represent the expected number of communities that is activated by the
set of nodes, \( S \subseteq B \). In former case, \( \vartheta(S) = \sigma(S) \), however in the latter case \( \nu(S) \) is directly proportional to \( \sigma(S) \). Remember, the objective function \( \sigma(S) \) represents the expected number of individuals influenced by the \( S \) nodes.

Equation (3.5) gives the redefined Shapley value formula. The worst case computational complexity of the redefined Shapley value is equivalent to naïve Shapley value computation when \( b = n \). This situation arises when there is no underlying community structure detected in the social network, a least probable event that is likely to occur in real world social networks. The average marginal gain of an individual \( i \) among the bridge nodes is defined in Equation (3.5)

\[
\frac{1}{b!} \sum_{\theta \in \Theta} \left[ \vartheta(S_i(\theta) \cup \{i\}) - \vartheta(S_i(\theta)) \right]
\]

where \( \Theta \) now represents set of all \( b! \) permutations of all bridge nodes \( B \), \( \Theta \) represents a single permutation in \( \Theta \) and \( S_i(\theta) \) represents the set of individuals that appear before node \( i \) in permutation \( \theta \). Nodes are then sorted in decreasing order based on their marginal gain value, and the top-K nodes in the list are identified as influential nodes.

### 3.4.5 Computational Complexity

Algorithm 3.1 takes \( O(E) \), where \( E \) is the number of edges in the social network. Algorithm 3.2 needs \( O(n\kappa) \), where \( \kappa \) represents the number of iterations in label propagation algorithm. Algorithm 3.3 that identifies bridge nodes across communities takes \( O(EM) \), where \( M \) is the number of communities. The running time of the Shapley value to compute the marginal contribution of each bridge node takes \( O(bE_b) \) times. \( E_b \) is the number of edges connecting bridge nodes, where \( E_b \ll E \). It takes \( O(b \log(b)) \) to rank the bridge nodes according to their Shapley values, which is negligibly very small. Taken together, the total worst case complexity is \( O(EM+n\kappa+bE_b) \).
3.5 PERFORMANCE ANALYSIS

The efficiency of the proposed algorithm is compared with the state-of-art SPIN algorithm presented by Ramasuri and Yadati (2011). Other algorithms available in the literature for finding top-K influential nodes have adopted greedy means, heuristic measures like degree centrality and probabilistic measures to find influential nodes. As the proposed approach and the SPIN algorithm are based on game theoretic concept in the context of information diffusion process, it would be more appropriate to compare these two algorithms. Experimental analyses are carried out in three real world datasets: (1) karate club (2) adjacency of nouns and (3) political books. These data sets are taken from the home page of Newman available at (http://www-personal.umich.edu/~mejn/netdata/). Brief descriptions of the datasets are given below.

1) Karate Club Dataset: This dataset describes the network of friendships between 34 members of a karate club in an US university. The dataset contains 34 nodes and 73 edges.

2) Political books Dataset: This dataset describes the selling pattern of books on US politics sold by the online bookseller Amazon.com. Nodes represent books about US politics and edges represent frequent co-purchasing of books by the same buyers, as indicated by the customers who bought this book also bought these other books feature on Amazon. This dataset contains 105 nodes and around 441 edges.

3) Adjacency of Nouns Dataset: This dataset contains the network of common adjective and noun adjacencies for the novel David Copperfield by Charles Dickens. Nodes represent the most commonly occurring adjectives and nouns
in the book. Edges connect any pair of words that occur in adjacent position in the text of the book. This dataset contains 112 nodes and around 400 edges.

The community structure inherent in these datasets and their nodes distribution are shown in Figure 3.2. Specifically, Figure 3.2(a) shows the three communities identified in karate club dataset and the number of nodes present in each community. Nodes within a community are clustered together based on their influence degree. Similar results for political books and adjacency of nouns datasets are shown in Figure 3.2(b) and Figure 3.3(c) respectively.

![Graphs showing community structure](image)

**Figure 3.2** Communities Identified using Linear Threshold Model
(a) Three Communities in Karate Club (b) Five Communities in Political Books and (c) Seven Communities in Adjacency of Nouns Datasets.
Here, the implementation details of the SPIN algorithm are presented that clearly explains the reason of better performance of the proposed approach. As working with different permutations of size $n$ is an intractable problem, the SPIN algorithm has chosen, to implement the problem based on perturbed permutations of size $q$, where $q$ is chosen appropriately based on the data set. A permutation is called as a perturbed permutation of size $q$ if it is a sequence of $q$ randomly selected nodes from the set of $n$ nodes, where $q < n$. In this study, $q$ nodes are selected from highly connected nodes that can better maximize the influence spread than the nodes that are sparsely connected. As the work in this chapter, deals with different permutations of $|b|$ the value of $q$ is also set to $|b|$. The Shapley values are then computed with various permutations of size $q$.

Figure 3.3 shows that the proposed approach gives better performance in terms of influence maximization than the SPIN algorithm. This is due to the fact that activating a single node in SPIN algorithm directly influences only its neighbors, whereas in the proposed work it activates the entire community to which the node belongs to, as communities are formed based on the nodes influence degree.
The next step in the proposed approach is to find top-$K$ nodes, the value of $K$ is set to $b$ in each dataset that will yield maximum influence spread. Figure 3.4 shows the running time of the proposed approach and the existing SPIN algorithm to find top 5 influential nodes in the political books dataset and top 7 nodes in adjacency nouns dataset. In the case of the political books dataset, to find the top 5 influential nodes, the SPIN algorithm takes 38 seconds whereas the proposed approach takes 5 seconds and for the adjacency of nouns dataset, to find the top 7 influential nodes, the SPIN algorithm takes 42 seconds whereas the proposed approach takes 6 seconds.
Figure 3.4  Running Time of the Proposed Algorithm and SPIN Algorithm for Top-5 Nodes in the Political Books and Top-7 Nodes in the Adjacency of Nouns Dataset

Figure 3.5  Shapley Value of Bridge Nodes for Various Datasets (a) Karate Club (b) Political Books and (c) Adjacency of Nouns

Figure 3.5 shows the Shapley value, the marginal gain achieved by activating the bridge nodes in various datasets. The empirical results obtained clearly reveals the advantages of the proposed approach over the SPIN
algorithm in terms of maximizing influence spread with much reduced time complexity.

3.6 SUMMARY

Active research has been into the area of finding influential nodes in social networks, for quite a longer time period as this domain has many important applications useful to mankind. Social network analysis metrics widely used in the literature to find influential nodes explores the underlying link structure that treats each individual as an independent entity. In reality, the behavior of the individual and the decision they make on a given issue depends on the society they belong to. Also, individuals tied together by positive relation can perform better in groups than in isolation. Game theoretic solution that studies the collective action of group of individuals is adopted in this study to find influential nodes in a social network. However, game theoretic solution spaces are not tractable and thus are found to be infeasible to study large social networks. The proposed approach attempts to reduce the complexity of the Shapley value solution space to be tractable based on the small-world characteristics and is thus scalable.