CHAPTER 1

INTRODUCTION

1.1 SCALE-FREE NETWORKS

The study of networks holds a prominent place in the field of scientific research. The widely explored kind of networks in the literature includes social networks, collaboration networks, citation networks, biological networks, information networks and many others. The advent of Web 2.0 and social networks have revolutionized the way people stay connected and share information. Traditionally, social networks have been studied in social sciences such as sociology, psychology, and business administration. The general features of these studies often consider the networks as graphs, whose nodes represent individuals and links represent the social interactions among these individuals. The increased availability of web based resources and on-line social networking sites have drew significant amount of attention from the research community to study the growth and evolution of these networks.

Among the networks, the studies on social, collaboration and citation network have a common objective to study the social links between the individuals, with varying degrees of intensity. Barabási and Albert (1999) demonstrated the existence of scale-free characteristics in these real life complex networks that is explained by the concept of preferential attachment. Scale-free network exhibits an important property in which a few nodes have a tremendous number of links to other nodes, whereas most of the nodes have
only very a few links. The popular nodes called hubs or influential nodes can have hundreds, thousands or even millions of links. In this sense, the network appears to have no scale, and thus called as scale-free networks. Following the preferential attachment process, new nodes arriving to the network prefer getting attached to the hub nodes. This leads to a power law degree distribution in the network that leads to rich-get-richer phenomenon.

An important current problem in all these domains is to locate influential nodes for information diffusion. Such nodes can effectively be utilized to maximize the influence spread in a society with less cost and effort. This problem gains a significant attention due to its wide applications such as viral marketing, epidemiology and identifying most prolific authors or documents in a research domain.

The main objective of this thesis is to define a mathematical as well as methodological framework for to identify and interpret influential nodes in a social, citation and collaboration networks. The findings obtained can contribute to a better understanding of network dynamics, its effect in network topology and the internal structure of links among individuals and communities.

1.2 MOTIVATION AND OVERVIEW

Individuals often get influenced by their acquaintances in the social network they belong to. Many factors can affect the strength of social influence. Social Impact Theory developed by Lantané (1981) identified three factors, which increase people’s likelihood to respond to social influence. They are the strength of the relationship between individuals, physical, temporal proximity and the number of individuals in the influencing group.
Watts (2003) noted that “We like to think of ourselves as individuals capable of making up our own minds about what we think are important, and how to live our lives. Individuals are to be seen as independent entities, their decisions are to be treated as originating from within, and the outcomes they experience are to be considered indicators of their innate qualities and talents. It's a nice story, implying as it does not only the theoretically attractive notion that individuals can be modeled as rationally optimizing agents, but also morally appealing message that each person is responsible for his or her own actions. However, there is a difference between holding someone accountable for their actions and believing the explanation for those actions is entirely self-contained. Whether we are aware of it or not, we rarely, if ever, make decisions completely independently and in isolation. Often we are conditioned by our circumstances, our particular life histories and our culture. We also cannot help but be influenced by the mesmerizing pool of universally available, often media-driven information in which we continually swim. In determining the kind of person that we are and the background picture against which our lives play out, these generic influences determine both the expertise and the preferences…”

Social impact theory and the above quote clearly highlight the fact that decision taken by an individual on a given issue depends on the behavior of their neighbors. This motivates the work carried out in this thesis in discovering a small set of influential individuals to perform certain task of information diffusion in a social, citation, collaboration networks and the details of these networks are discussed in the subsequent sections.

1.2.1 Social Networks

Social networks provide great opportunities for social connections for learning, individual entertainment and enhancement in a wide variety of forms. Study on social networks gain significance for the role they play in
sharing and disseminating information at a massive scale. Information can travel widely through a social network via word-of-mouth exchanges between friends in the network. Recently, significant amount of research efforts have been done in deriving models that can explain how information propagates, contagious disease spreads, and even terrorist organizations operate. In a classic study, Coleman et al (1966) showed how doctors’ willingness to prescribe a new antibiotic tetracycline diffused through professional contacts. Several deterministic mathematical models proposed by Granovetter (1978), Watts (2002) and stochastic model introduced by Brock and Durlauf (2001) describes the information diffusion in a social network, where individuals decide whether or not to adopt a certain behavior in part, or in whole, based on the previous decisions of others.

In these models, an individual can influence others to adopt a technology or buying a product, and they can in turn influence several others in a cascading fashion. This leads to the involvement of maximum number of people to participate in the social network, which is called as social influence. In the process of social influence, the number of individuals initially selected plays a crucial role for maximizing the influence. The work proposed in this thesis analyzes effective approaches to locate such influential nodes.

The following section gives a brief description about the significance of finding such influential nodes in the citation and collaboration networks.

1.2.2 Citation and Collaboration Networks

Citation and collaboration networks are widely used to study the scientific communication and knowledge flows that occur among the individuals (researchers) present in those networks. A systematic review of past literature is an essential feature and crucial endeavor for any academic
research. Hart (1998) pointed out that it is initially required for any researcher to broadly read on any topic to know the advances of a field. Researchers may be able to redefine and narrow down their scope of reading, after they clearly identify the research problem even though plethora of information sources appears to be pertinent. A practical tip to effectively manage the literature review and identifying when the literature review is complete is given by Levy and Ellis (2006). Bandara et al (2011) have analyzed the use of qualitative tool that can assist with the management of large amounts of literature information across many different phases of a long term research study.

Sir Isaac Newton (1676) noted “if I have seen a little further it is by standing on the shoulders of giants”. Clearly, researchers who are able to take advantage of knowledge and discoveries of other researchers, can better understand the advancements in a given problem domain. Therefore, quality research literature from leading, peer-reviewed journals and researchers should serve as the major base of literature review, as it provides sufficient theoretical background as well as leads for additional references on the specific subject matter.

Webster and Watson (2002) proposed the idea of backward and forward searching that respectively involves searching for the reference or citing articles and authors. The idea of backward references search refers to initially finding relevant articles from keyword search, followed by reviewing the references of the retrieved articles. Levy and Ellis (2006) exemplified backward references search by pointing out that study on the construct of computer self efficacy, by keywords search retrieves Compeau Higgins’s (1995) pioneering article. Conducting a backward references search for computer self efficacy, means to retrieve and review all the relevant references of Compeau Higgins’s (1995) article such as Bandura (1977) or
Barling (1983) articles. Doing such backward reference search provides researchers with the ability to learn more about the origins of the construct, theory, or model under study. On the other hand, backward authors search refers to reviewing what the authors have published prior to the article. Querying an author’s prior work may yield useful information. For example, users would come across Doll and Torkzadeh (1991) article in their study on the construct of end user computer satisfaction (EUCS). By performing backward author search one can locate other articles that these authors have analyzed that have led them to the development of the EUCS construct.

Forward reference search refers to reviewing additional articles that have cited the article. For example, a forward reference search for Compeau and Higgins’s (1995) article can be accomplished by conducting a search for all articles that include either Compeau or Higgins in their citations. Forward author search refers to reviewing the publications of other authors following this article. This would help to track subsequent research progress carried out following the line of investigation done by Compeau and Higgins’s.

Performing such searches enable the researcher to expand their knowledge by identifying improvements or new findings related to the phenomenon under study. Thus, identifying influential researchers and documents gain foremost importance in the study of a research domain. Further, the influential nodes in the study of citation and collaboration networks are expected to be active in a given time period. This issue gains significance as the documents published in recent time period and active researchers in a problem domain would present current trends and further research directions. However, due to preferential attachment, nodes added to the network at an earlier time can receive more influence than the nodes added later. This thesis provides effective solutions for locating active as well influential nodes.
1.2.3 Limitations

The field of social network theory primarily focused on measuring the structural properties of naturally occurring networks. That is, social network theory models the pattern of links between nodes in a network rather than attempting to model the behavior of the individuals that these nodes represent. Most of the work in this field involved the study of statistical properties of networks such as degree of distribution of the nodes, the average number of links the nodes maintain, the density of the edges, and the diameter of the entire network. These types of statistics are often measured in real networks like phone call graphs, e-mail networks, on-line communities, and co-authorship networks (Kossinets and Watts (2006), Backstrom et al (2006)). Once these types of statistics are found in a variety of real world networks, generative models are created that output networks with similar statistical properties. Also, these networks grow by adding nodes and edges via stochastic dynamics. For example, in the case of preferential attachment model, a new node gets attached to an existing node with probability proportional to the existing node’s degree. Most studies on network growth models capture only its structure but not the dynamics of interaction that occur over these networks. The field of game theory overcomes these problems, as it provides a rich suite of mathematical models for analyzing strategic interaction among autonomous, intelligent and rational individuals (or players). Recently, there have been several efforts in following a game theoretic approach to social network modeling (Ramasuri and Narahari (2011), Aadithya et al (2010)). However, studies in the literature based on the game theoretic approach to solve influential node problem in the social networks rarely studied the influence of small-world characteristics. In this thesis attempts are made to enhance the performance of the state-of-the-art game theoretic algorithms, by proposing a framework that exploits the
small-world characteristics such as community structure and the weak tie properties of social networks.

Finding influential authors, journals or articles in scholarly networks (citation and collaboration) is measured based on various citation analyses. Citations are understood to serve as carriers of authority and correspond to different endorsements. Scientific collaboration and endorsement are well-established research topics utilizing three kinds of methods presented by Stasa (2010): (1) qualitative methods (e.g., surveys/questionnaires, interviews, or observations), (2) bibliometric methods (e.g., publication counting, citation counting, or co-citation analysis) and (3) complex network methods (e.g., shortest path, centralities, network parameters, or PageRank/HITS). These studies either provide quantitative analytical results about the global network features or rankings of individual nodes, or qualitative content analysis of survey/observation results. Some ranking algorithms (e.g., PageRank, HITS) take the whole network as a graph and utilize the probability propagation of random surfers based on how many times a node is linked by other nodes, as well as the number of times a node is linked by other important nodes. Mainstream network analysis research uses mathematical models and graph theory to analyze and categorize nodes based on network connectivity properties and summarizes their distribution patterns. Most network analysis does not consider topical features of nodes (individuals) that can effectively capture their evolving knowledge, subsequently failing to locate active nodes.

Motivated by the above observation, this thesis delves to: (1) analyze game theoretic models than can effectively capture the influence of small-world characteristics in analyzing the strategic nature of individuals in the social network and (2) present novel solutions to track the evolving knowledge of individuals (researchers) in the scientific community.
The terms influential nodes, authors, researchers and experts are used interchangeably throughout the thesis. Similarly, the terms research article, paper and document are used interchangeably in the context of citation and collaboration networks. Also, the term scholarly network is used to refer both citation and collaboration networks.

1.3 RESEARCH CONTRIBUTION

The work reported in this thesis focused on developing effective solutions to an interesting and key problem of finding influential nodes, under various solution spaces in social and scholarly networks, and the main contributions are given in the subsequent paragraphs.

Co-operative game theoretic approach that provides rich mathematical model to study the interaction among individuals is the first solution space considered in this thesis to identify top-K influential nodes in the social networks. Top-K nodes problem is a variant of influential node selection problem where it is required to identify $K$ nodes that maximizes influence spread in a network. Top-K nodes problem studied by Kempe et al (2003) proved that the optimization problem of selecting the most influential nodes is NP-hard, and derived the first greedy algorithm with provable approximation guarantees. Ramasuri and Narahari (2011) have studied the top-K nodes problem as cooperative game and the coalitional strength among players are computed using Shapley values, the solution space of cooperative game theory.

Sampling based approach is widely used in the literature to calculate Shapley values approximately, as exact Shapley value computation is intractable. The effectiveness of the solution obtained is thus limited by the kind of samples chosen. Also, the computational complexity of the existing game theoretic algorithms is high and does not provide scalable solutions for
large social networks. The proposed work thrives to reduce the computational complexity by taking into account the small-world properties inherent in social networks. Small-world networks such as social networks exhibit high clustering coefficient and sparse node connectivity. High clustering coefficient indicates presence of communities and in sparse node connectivity most of the nodes are not neighbors of one another but they can be reached from every other by a small number of hops or steps.

An effective method is proposed in the thesis, for locating influential nodes in a social network. Initially, communities are detected based on the influences between nodes and the bridge nodes that enable communication between nodes belonging to different communities. The ability of the bridge nodes to maximize the influence spread in the whole network compared to other nodes is studied by Chwe (1999). Top-K influential nodes among the bridge nodes are then returned based on their marginal gains.

Co-citation analysis is another solution space explored in this dissertation to locate influential authors in a research domain. Co-citation analysis introduced by Small (1973) examines the intellectual structure of many disciplines by analyzing a set of items (authors, documents, journals, etc.) that represent a research area and their relationships using co-citation counts. Co-citation count between two items is defined as the number of times they are cited together by later studies in a research domain. The proposed work in the thesis analyzes the co-occurrence of authors, the method that is widely studied in the literature.

Author Co-citation Analysis (ACA) initially proposed by White and Griffith (1981) is a tool widely used to identify underlying structure in a scientific discipline. The quality of author clusters produced in the existing works on ACA depends on the co-citation count that does not include
temporal dimension in its analysis. Thus, authors who published their work earlier get high co-citation count due to preferential attachment than the contemporary authors. However, users would be interested in identifying active and influential authors with up-to-date knowledge in a problem domain. Tracking such potential authors is further beneficial in scenarios such as identifying research collaborators, reviewers of technical papers, experts for presenting invited talks in conferences, seminars, etc.

A novel co-citation counting technique is proposed in this thesis to study the underlying dynamics in the citation network. The co-citation count is suitably modified by an aging parameter that includes the current time period and publication time of the research documents. Further, it is extended to include the number of times authors of two different research papers cite together earlier studies in a problem domain. The effectiveness of the proposed approach is evaluated with ACM Computing Classification System (CCS) benchmark.

PageRank, a popular link analysis algorithm developed by Brin and Page (1998) is the next solution space, which is proved to be effective in finding influential nodes in the citation network. These algorithms measure the importance of research publications from another perspective such as the prestige of the one who actually cited the papers. PageRank algorithm that analyzes the link structure of web is then extended to citation analysis by Ma et al (2008). They have compared the consistency and disparity between PageRank and citation count, and found that the PageRank and citation count are strongly correlated. As PageRank incorporates the importance of citing papers to a specific cited paper, it excavates several important papers that may suffer from low citations.

The need for topic-specific link analysis arises as the work done by Chakrabarti et al (2002) demonstrates that PageRank is not sensitive to the
page topic. As a result, pages considered important in some subject domains with a few incoming links may not be ranked high and that can be avoided by making PageRank topic-sensitive. A few studies proposed by Chakrabarti et al (2002) and Haveliwala (2003) used fixed topic taxonomies such as Yahoo! or the Open Directory Project (ODP) to understand the structure of content-based clusters and communities. TwitterRank is proposed by Weng et al (2010) to measure the influence of a twitter, taking both the topic similarity between users and the link structure into account.

Fixed size topic taxonomies impose a severe constraint as new topics are likely to emerge in dynamic environment like a research domain in the citation network. Also, the returned topic-specific contents may not be recent ones, as temporal dimension is not included in the analysis. The thesis work proposes novel solutions to extract latent topic information automatically from the underlying text corpus in the citation network using Latent Dirichlet Allocation model, an unsupervised machine learning technique proposed by Blei et al (2003). Topic-specific PageRanks of documents (implicitly represent its authors) are computed, and contemporary authors in a given problem domain are retrieved by suggesting an appropriate aging factor.

Referral system provides a natural solution space that mimics human interaction to find experts in a problem domain. Kautz et al (1997) approached the expert finding problem from a different perspective. Their work is based on the observation that expert finding is an inherently complex process. Hence, they contend that the best way of finding an expert is through referral chaining whereby a seeker finds the needed expert through referral by colleagues. A referral model in which a node’s behavior of looking for an expert is adjusted by the individual’s awareness of the potential expertise of their contacts is proposed by Xu and Victor (2010).
Existing expert finding system through referrals mostly analyzes a static social environment. As the expertise of an individual changes with time, static social networks may not effectively locate experts at a given time instance. Thus, the proposed work in the thesis attempts to locate influential nodes through referrals in time-evolving social networks.

The time-evolving social network is viewed as a sequence of graphs taken at different time slices. In this scenario, tracking the evolving knowledge of an expert is inevitable, which would otherwise lead to global expertise re-computation. A dynamic knowledge model based on an aggregation operator is proposed to effectively capture the knowledge evolution of an expert across various time slices. Each individual is also associated with an expert cache that can effectively find more influential nodes with less number of referrals.

The notion of profiles that provides another solution space for locating influential nodes in a problem domain is studied by Balog and Rijke (2007). The profile of a candidate contains documentary evidence representing the expertise of this candidate (e.g., documents, articles, project reports, or e-mails authored by or that contain some sort of identification of the candidate). The profiles of the candidates can be used to rank them according to their expertise in a given topic.

A new dimension of expert search whether some documents in a candidate’s profile serves as better indicator for expert search is investigated by Macdonald et al (2008). It is pointed out by Smirnova and Balog (2011) that in a real-world organizational setting, the notion of the best expert depends on the individual user performing the search and an user-oriented model is proposed that incorporates user-dependent factors such as preferred contact time of the user and the knowledge gained so far in the topic.
Existing works in this solution space mostly focused on locating experts in a single research domain. However, in reality, many research works carried out are interdisciplinary in nature. Solution to find experts with adequate knowledge in more than one problem domain can effectively meet such real-world requirements. A modified vector space method for document representation is proposed in this thesis. The data structure, Telescopic Vector (TV) trees proposed by Lin et al (1994) for indexing and efficiently processing high dimensional data space is used to represent and retrieve documents. The retrieved documents relevant to a problem domain, provides documentary evidence to compute author’s score. The relative expertise of the author is estimated by computing the ratio of author score with the combined score of other authors associated with the retrieved documents. Influential authors with multiple expertise are then identified based on their expertise score in the given problem domain.

1.4 RELATED TERMINOLOGIES

This section gives an overview of various terminologies that are used in the thesis.

**Graph**: A network with $n$ number of nodes connected by a set of $e$ links is represented by the graph $G (V, E)$ where $V$ and $E$ represent the set of nodes and set of edges respectively such that $|V|= n$ and $|E|=e$.

**Graph Diameter**: The diameter of a graph is the maximum eccentricity of any vertex in the graph. That is, it is the greatest distance between any pair of vertices.

**Betweenness**: Vertices that occur on many shortest paths between other vertices in a graph $G$ have higher betweenness than those that do not.
**Neighborhood:** The neighborhood of a vertex \( v \) in a graph \( G \) is the induced sub graph of \( G \) consisting of all vertices adjacent to \( v \) and all the edges connected to \( v \). The neighborhood \( N_i \) for a vertex \( v \) is its immediately connected neighbors and is defined in Equation (1.1).

\[
N_i = \{v_j: e_{ij} \in E \land e_{ji} \in E\}
\]  
(1.1)

**Path in a Graph:** It is a sequence of vertices such that from each of its vertices there is an edge to the next vertex in the sequence.

**Path length:** It is the number of edges that the path uses counting multiple edges multiple times.

**Weighted Graph:** It associates a value (weight) with every edge in the graph. The weight of a path in a weighted graph is the sum of the weights of the traversed edges.

**Clique:** A clique in an undirected graph \( G = (V, E) \) is a subset \( C \) of the vertex set \( V \), such that for every two vertices in \( C \), there exists an edge connecting the two.

**Clustering:** It is an unsupervised learning technique by which the data set can be grouped using some distance metric between the data elements. There are two main types of clustering techniques, hierarchical and non hierarchical used as classical data mining techniques.

**Clustering coefficient:** It is a measure of degree to which nodes in a graph tend to cluster together. Global and local are the two types of clustering coefficient. The global clustering coefficient gives an overall indication of the clustering in the network, whereas the local gives an indication of the embeddedness of single nodes.
**Connected component:** It is a maximal connected sub graph in which every vertex \( x \) is reachable to every other vertex \( y \). It can be defined as \( G (V, E) \) where \( V \) represents the set of nodes in the sub graph connected by the set of edges \( E \).

**Cut vertex:** It is a vertex in the graph whose removal disconnects the graph into a set of connected components. The cut vertices can be identified by performing depth first search on the graph in \( O(n+e) \) where \( n \) is the number of nodes and \( e \) is the number of edges in the graph.

**Degree centrality:** It is defined as the number of links incident upon a node. Degree is often interpreted in terms of influence of the node in spreading information.

**Kendall's rank correlation:** In statistics, the Kendall rank correlation coefficient, commonly referred to as Kendall’s Tau \((\tau)\) coefficient, which is a statistic used to measure the association between two measured quantities. A \( \tau \) test is a non-parametric hypothesis test that uses the coefficient to test for statistical dependence. Let \((x_1, y_1), (x_2, y_2),..., (x_n, y_n)\) be a set of joint observations from two random variables \( X \) and \( Y \) respectively, such that all the values of \( \{x_i|1 \leq i \leq n\} \) and \( \{y_i|1 \leq i \leq n\} \) are unique. Any pair of observations \((x_i, y_i)\) and \((x_j, y_j)\) are said to be concordant if the ranks for both elements agree; that is, either \( x_i > x_j \) and \( y_i > y_j \) or \( x_i < x_j \) and \( y_i < y_j \). They are said to be discordant, when it satisfies either \( x_i > x_j \) and \( y_i < y_j \) or \( x_i < x_j \) and \( y_i > y_j \). If \( x_i = x_j \) and \( y_i = y_j \), the pair is neither concordant nor discordant. The Kendall coefficient is defined in Equation (1.2).

\[
\tau = \frac{\text{No. of concordant pairs} - \text{No. of discordant pairs}}{\frac{1}{2} n(n-1)} \tag{1.2}
\]
**Preferential Attachment:** A class of process in which some quantity, typically some form of wealth or credit, is distributed among a number of individuals according to how much they already have, so that those who are already wealthy receive more than those who are not.

**Power Law:** It is a special kind of mathematical relationship between two quantities. When the frequency of an event varies as a power of some attribute of that event (e.g. its size), the frequency is said to follow a power law.

**Scale-free network:** A network whose degree distribution follows a power law, at least asymptotically. That is, the fraction $P(k)$ of nodes in the network having $k$ connections to other nodes goes for large value of $k$ as $P(k) \sim c k^{-\gamma}$, where $c$ is a normalization constant and $\gamma$ is a parameter whose value is typically in the range $2 < \gamma < 3$. Many networks are conjectured to be scale-free, including social, citation and collaboration networks, World Wide Web links and biological networks.

**Social network analysis:** Views social relationships in terms of network theory consisting of nodes and ties (also called edges, links, or connections). Nodes are the individuals within the networks, and ties are the relationships between them. The resulting graph-based structures are often very complex. There can be many kinds of ties between the nodes. Research in a number of academic fields has shown that social networks operate on many levels, from families up to the level of nations, and play a critical role in determining the way problems are solved, organizations are run, and the degree to which individuals succeed in achieving their goals.

**Spearman's correlation coefficient:** In statistics, Spearman's rank correlation coefficient or Spearman's rho ($\rho$) is a non-parametric measure of statistical dependence between two variables. It evaluates how well the relationship between two variables can be described using a monotonic
function. The \( n \) raw scores \( X_i \) and \( Y_i \) are converted to ranks \( x_i \) and \( y_i \). Differences \( d_i = x_i - y_i \) between the ranks of each observation on the two variables are calculated, and the value of \( \rho \) is given in Equation (1.3).

\[
\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{1.3}
\]

### 1.5 ORGANIZATION OF THE THESIS

The rest of the thesis is organized as follows. Chapter 2 provides a survey of related research works in the literature. Chapter 3 defines a game theoretic approach to locate influential nodes in the social networks. Chapter 4 presents the proposed Time-based Author Co-citation Analysis (TACA) to locate active and influential researchers in a citation network. A novel co-citation counting technique that studies the underlying dynamics in the network is explained in this chapter. Chapter 5 deals with the proposed method of finding active and influential researchers using link based analysis approach. Topic-specific PageRank and an aging factor computation methodology to retrieve contemporary and influential authors (documents) in a given topic are discussed. Chapter 6 explains the referral systems that naturally capture the manner in which people help each other in finding trustworthy experts. A dynamic knowledge model that studies the evolving knowledge of an expert through learning in a time-evolving social network is proposed in this chapter. Chapter 7 defines a content-based analysis approach to locate researchers with multiple expertise. A modified vector space is proposed to represent documents and the Telescopic Vector (TV) tree data structure that efficiently process this high dimensional vector are discussed in this chapter. Chapter 8 presents the concluding remarks and the scope for further enhancements.