SUPERVISED LEARNING BASED APPROACH TO LINK PREDICTION UTILIZING SOCIAL BALANCE THEORY

Dissertation submitted to Jawaharlal Nehru University in partial fulfillment of the requirements for the award of the Degree of

MASTER OF TECHNOLOGY
In
COMPUTER SCIENCE AND TECHNOLOGY

By
ARTI PATIDAR

Under the Guidance of
PROF. K. K. BHARADWAJ

SCHOOL OF COMPUTER & SYSTEMS SCIENCES
JAWAHARLAL NEHRU UNIVERSITY
NEW DELHI –110067
JULY 2012

JAWAHARLAL NEHRU UNIVERSITY
NEW DELHI - 110067

CERTIFICATE

This is to certify that the dissertation entitled "**Supervised Learning Based Approach To Link Prediction Utilizing Social Balance Theory**", being submitted by **Arti Patidar** to the School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, in partial fulfillment of the requirement for the award of the Degree of **Master of Technology in Computer Science and Technology**, is a bona fide work carried out by her under the guidance and supervision of **Prof. K. K. Bharadwaj**.

The matter embodied in the dissertation has not been submitted for the award of any other Degree or Diploma.

Patidar

Arti Patidar

(Student)

25-07-2012

Prof. K. K. Bharadwaj

(Supervisor)

Prof. Karmeshu
Dean, SC&SS
Jawaharlal Nehru University
New Delhi-67
India.

प्रोफेसर कर्मेशु/Professor Karmeshu
डीन / Dean
स्कूल ऑफ कम्प्यूटर एवं पद्धति विज्ञान संस्थान
School of Computer and Systems Sciences
जवाहरलाल नेहरू विश्वविद्यालय
Jawaharlal Nehru University
नई दिल्ली / New Delhi-110067

Acknowledgement

It gives me immense pleasure in expressing my gratitude to all the people who have made this dissertation possible. First and foremost I would like to thank my guide Prof. K.K. Bharadwaj, who gave me great intellectual freedom to pursue my interest and provided encouragement and guidance throughout the work. His extraordinary interest and inspiring attitude helped me in solving all the hurdles that cropped up at various stages. Working with him has been an excellent learning experience. The dissertation would not have been possible without his encouragement, support and patient guidance.

I also gratefully acknowledge Dean, Prof. Karmeshu for providing the necessary infrastructure to carry out this dissertation. I take immense pleasure in thanking to all faculty members for imparting knowledge to me and to all staff members for providing me necessary facilities in due course of this work.

No words are enough for my parents who are always the source of inspiration for all my achievements. My heartiest thanks are due to my siblings, Rohit, Kanchan, Khushboo and Pooja for being very supportive and encouraging throughout the process. Of course, a very special thanks to all seniors and my friends, who shared their knowledge with me and for their timely support throughout tenure. Specially thank to my classmate Ghyani Umesh, my labmates Mr. Vibhor Kant, Ms. Pragya Dwivedi and Ms. Vinti Agrawal as they accompanied and suggested me a lot during my research work. It is a great privilege to have had the opportunity to work with them.

In the end, I would like to take this opportunity to express my gracious gratitude towards all those people who have in various ways helped me in making this work on éclat.

ARTI PATIDAR

Abstract

Due to explosive growth of social network sites like MySpace, Facebook, Linkedin, Flickr, Orkut etc., social network analysis is manifested as a separate field to understand, manage and study these virtual online social networks and became much explored and studied area from the last decade. Social networks are usually modeled using directed graphs, where an edge between two nodes represents a relationship between two individuals. A basic computational problem underlying social-network evolution is *link prediction problem* which aims at estimating the likelihood of the existence of a link between two nodes of the network. Sometimes social interactions also involve negative relationships besides the positive one. Positive links indicate friendship, support, or approval and negative link signify enmity, disapproval of others, or distrust of the opinions of others. In our work, we are predicting positive and negative links to recommend 'friends' and 'foes' to individuals in online social networks.

Social networks are the patterns of contact that are created by the flow of messages among communicators through time and space. The concept of message should be understood here in its broadest sense to refer to data, information, knowledge, images, symbols, diseases, infections, signals and any other symbolic forms that can move from one point in a network to another or can be co-created by network members. In real life these networks are formed because of some motives, intentions or say reasons, which collectively form an important part of *social science theories*. To bring virtual online social networks more closer to real social networks the computer scientist are trying to incorporate these theories in online social network. One such theory which we are using in this dissertation work is *Social Balance Theory*.

In this work, we propose a social recommender system for social network sites, which not only has friends but also foes to recommend. We first employ inductive learning heuristics to extract rules from the features patterns of existing friends & foes of an individual. Thereafter, we recommend positive and negative links, using the extracted rules, to an individual avoiding possible social imbalance in the extended friends & foes network of an individual based on social balance theory.

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Since their introduction, social network sites (SNSs) such as MySpace, Facebook, Linkedin and Orkut have attracted millions of users, many of whom have integrated these sites into their daily practices. Today People are more willing to seek their interests from the Web. As of this writing, there are hundreds of SNSs, with various technological affordances, supporting a wide range of interests and practices. For example, movie fans could rent DVDs according to the recommendations from Netflix. Flickr provides a platform for Web users to seek and post favorite photos. Facebook can connect people with their communities sharing similar interests. Web users could obtain and contribute knowledge through Wikipedia.

Scholars from disparate fields have examined SNSs in order to understand the practices, implications, culture, and meaning of the sites, as well as users' engagement with them. While the key technological features of all these SNSs are fairly consistent, the cultures that emerge around SNSs are varied. Most sites support the maintenance of pre-existing social networks, but others help strangers to connect based on shared interests, political views, or activities. Some sites cater to diverse audiences, while others attract people based on common language or shared racial, religious, or nationality based identities. Sites also vary in the extent to which they incorporate new information and communication tools, such as mobile connectivity, blogging and photo/video-sharing.

The following sections give a brief overview of online social networks, signed networks, link mining, social recommender systems and organization of thesis.

1.1 **Social Networks**

A **social network** is a social structure made up of individuals (or organizations) called "nodes", which are tied (connected) by one or more specific types of interdependency, such as friendship, kinship, common interest, financial exchange, dislike, sexual relationships, or relationships of beliefs, knowledge or prestige. Natural examples of social networks include the set of all scientists in a particular discipline, with edges joining pairs who have coauthored articles; the set of all employees in a large company, with edges joining pairs working on a common project; or a collection of business leaders, with edges joining pairs who have served together on a corporate board of directors [Liben and Kleinberg, 2007].

A **social network site** can be defined as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site [Boyd and Ellison, 2007]. At the commencement of chapter we already mentioned some of the most popular SNSs, some of the more SNSs are Twitter, Youtube, Friendster, Hi5, Yahoo! 360, Flicker and so on.

Some authors used the same above mentioned definition but referred SNS as **web based social network (WBSN)** [Golbeck, 2008] and **online social network (OSN)** [Mislove, 2009]. In this dissertation, we are using all the three terms interchangeably. To keep any website in the category of SNS it must fulfil some more criteria besides what is given in definition above. These criteria are as follows:

1. It is accessible over the web with a web browser.
2. Users must explicitly state their relationship with other people.
3. The system must have explicit built-in support for users making these connections.
4. Relationships must be visible and browsable.

Unlike the traditional Web, which is largely organized by content, online social networks embody users as first-class entities. Users join a network, publish their own content, and create links to other users in the network called “friends”. This basic user-to-user link structure facilitates online interaction by providing a mechanism for organizing both real-world and virtual contacts, for finding other users with similar interests, and for locating content and knowledge that has been contributed or endorsed by “friends”.

The extreme popularity and rapid growth of these online social networks represents a unique opportunity to study, understand, and leverage their properties. A large contribution of specific journals and conferences has been made towards the rapid development of network theory and methods. The growth on **Social Network Analysis (SNA)** as an academic field has emerged and coincided with an explosive interest of people in social networks.

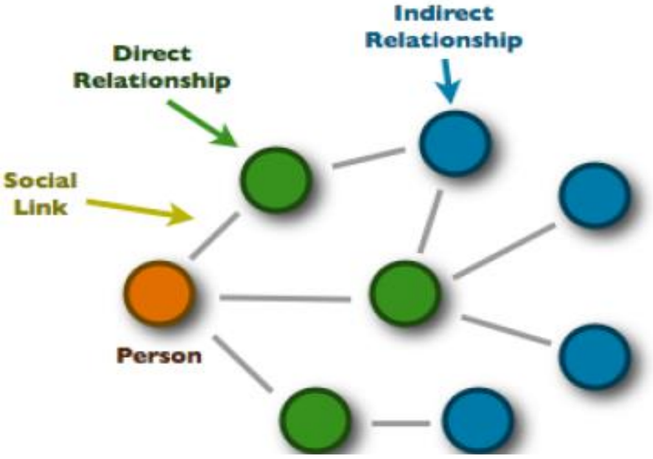

Figure 1.1: Social network representation as graph

1.1.1 Social Network Analysis

Social network analysis has emerged as a key technique in modern sociology. It has also gained a significant following in anthropology, biology, communication studies, economics, geography, information science, organizational studies, social psychology, sociolinguistics etc. and has become a popular topic of speculation and study.

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 individual actors within the networks and ties are the relationships between the actors. 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.

1.1.2 Evolution of Online Social Networks

We now give a brief history of online social networks. The site Classmates.com is regarded as the first web site that allowed users to connect to other users. It began in1995 as a site for users to reconnect with previous classmates. However, Classmates.com did not allow users to create links to other users; rather, it allowed users to link to each other only via schools they had attended. In 1997, the site SixDegrees.com was created, which was the first social networking site that allowed users to create links directly to other users. As such, SixDegrees.com is the first site that meets the definition of an online social network from above. While SixDegrees attracted millions of users, it failed to become a sustainable business and in 2000, the service closed. Shortly after its launch in 1999, LiveJournal listed one-directional connections on user pages. On LiveJournal, people mark others as *friends* to follow their journals and manage privacy settings. The Korean virtual worlds site Cyworld was started in 1999 and added SNS features in 2001.

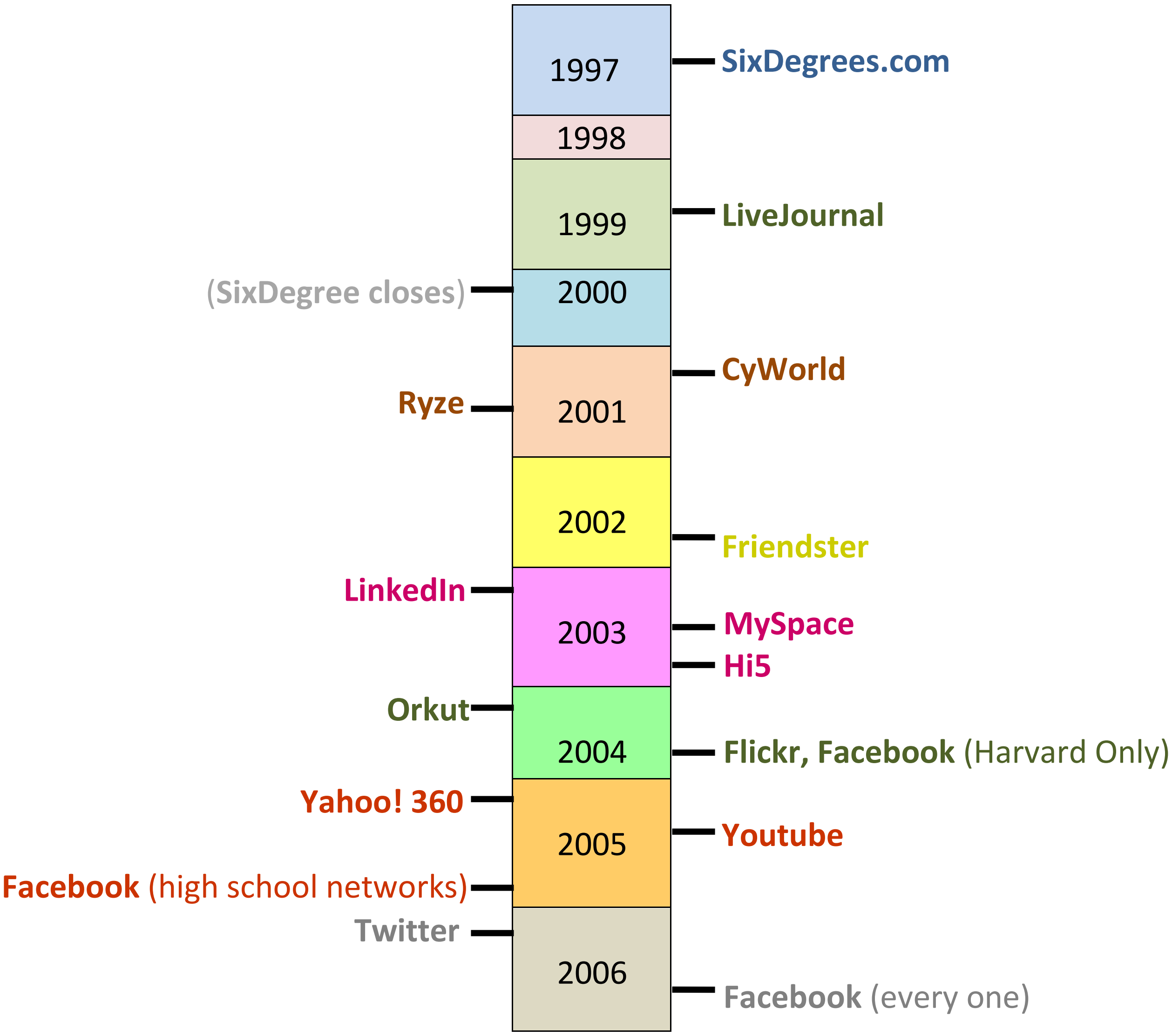

Figure 1.2: Timeline of the launch dates of many major SNSs

The next wave of SNSs began when Ryze.com was launched in 2001 to help people leverage their business networks. Ryze was first introduced to primarily members of the San Francisco business and technology community, including the entrepreneurs and investors behind many future SNSs.

Friendster, MySpace and Facebook were the next three key SNSs that shaped the business, cultural and research landscape. Friendster launched in 2002 as a social complement to Ryze. Friendster was focused on allowing friends-of-friends to meet, beginning as a rival to the online dating site Match.com. In 2003, MySpace was created as an alternative to Friendster and the others. MySpace allowed users to heavily customize the appearance of their profile, which proved very popular with users, causing MySpace to quickly become the largest online social network. Other, similar sites created in the same timeframe include LinkedIn and Hi5.

Orkut was started in early 2004 and very soon became as popular as Myspace. Before gaining popularity in India it became famous in Brazil. Facebook also began in early 2004 as a Harvard-only SNS. To join, a user had to have a harvard.edu email address. As Facebook began supporting other schools, those users were also required to have university email addresses associated with those institutions, a requirement that kept the site relatively closed and contributed to users' perceptions of the site as an intimate, private community. Unlike other SNSs, Facebook users are unable to make their full profiles public to all users. Beginning in September 2005, Facebook expanded to include high school students, professionals inside corporate networks and eventually, in 2006 it included everyone.

With the rise in popularity of online social networks, many other types of sites began to include social networking features. Examples include multimedia content sharing sites (Flickr and YouTube), blogging sites (LiveJournal and BlogSpot), professional networking sites (LinkedIn and Ryze), and news aggregation sites (Digg, Reddit and del.icio.us). All of these sites have different goals but employ the common strategy of exploiting the social network to improve their sites. The list above is not meant to be exhaustive, as new sites are being created regularly. For a more complete history and analysis of the evolution of online social networks, readers are referred to the numerous papers by Boyd [Boyd, 2004; Boyd, 2006; Boyd and Ellison, 2007].

The introduction of SNS features has introduced a new organizational framework for online communities, and with it, a vibrant new research context. As discussed earlier social networks are usually represented as a unipartite directed graph, where vertices represent objects and edges represent relationships between those objects. Hence we extensively use graph theory to state and measure the properties of the network. Misolve had described the structure of OSN in great detail in his PhD thesis [Mislove, 2009].

1.1.3 Properties of Online Social Networks

There are few properties that seem to be common to many networks: The small-world property, power-law degree distributions, and network transitivity. Small world effect is the finding that the average distance between vertices in a network is short, usually scaling logarithmically with the total number of vertices. The degree of a vertex in a network is the number of other vertices to which it is connected, and one finds that there are typically many vertices in a network with low degree and a small number with high degree, the precise distribution follow a power-law or exponential form. Analysis of large-scale growth data shows that new links are created and received by users in direct proportion to their current number of links. There are various measures in OSN which is helpful in describing properties of networks. Some of these are given below:

- Radius and diameter
- Degree distribution
- Joint degree distribution
- Scale-free behaviour
- Assortativity
- Clustering coefficient
- betweenness centrality
- Connected components

1.2 Link Mining

Link mining is a newly emerging research area that is at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining [Getoor and Diehl, 2005]. Link mining is an instance of multi-relational data mining; however, we use the term *link mining* to put an additional emphasis on the links—moving them up to first-class citizens in the data analysis endeavour.

Link mining encompasses a range of tasks including descriptive and predictive modelling. Both classification and clustering in linked relational domains require new data mining algorithms. But with the introduction of links, new tasks also come to light. Examples include predicting the numbers of links, predicting the type of link between two objects, inferring the existence of a link, inferring the identity of an object, finding co-references, and discovering subgraph patterns. We classify these tasks hierarchically in Section 1.2.2.

1.2.1 Data Representation

While data representation and feature selection are significant issues for traditional machine learning algorithms, data representation for linked data is even more complex. Consider a simple example of a social network describing actors and their participation in events. Such social networks are commonly called affiliation networks [Wasserman and Faust, 1994] and are easily represented by three tables representing the actors, the events, and the participation relationships. Even this simple structure can be represented as several distinct graphs. The most natural representation is a bipartite graph, with a set of actor nodes, a set of event nodes, and edges that represent an actor's participation in an event. Other representations may enable different insights and analysis. For example, we may construct a network in which the actors are nodes and edges correspond to actors who have participated in an event together. This representation allows us to perform a more actor-centric analysis. Alternatively, we may represent these relations as a graph in which the events are nodes, and events are linked if they have an actor in common. This representation may allow us to more easily see connections between events.

This flexibility in the representation of a graph arises from a basic graph representation duality. This duality is illustrated by the following simple example: Consider a data set represented as a simple G(O, L), where O is the set of objects (i.e., the nodes or vertices) and L is the set of links (i.e., the edges or hyperedges). The graph G(O, L) can be transformed into a new graph G'(O', L'), in which the links li, lj in G are objects in G' and there exists a link between $O_i , O_j \in O'$ if and only if li and lj share an object in G. This basic graph duality illustrates one kind of simple data representation transformation. The representation chosen can have a significant impact on the quality of the statistical inferences that can be made. Therefore, the choice of an appropriate representation is actually an important issue in effective link mining, and is often more complex than in the case where we have IID data instances.

1.2.2 Taxonomy of Common Link Mining Tasks

As mentioned before, link mining puts a new twist on some classic data mining tasks, and also poses new problems. Here we provide a list of possible tasks defined by [Getoor and Diehl, 2005]. These tasks are classified according to the part of the graph to be mined using mainly information attached with the links.

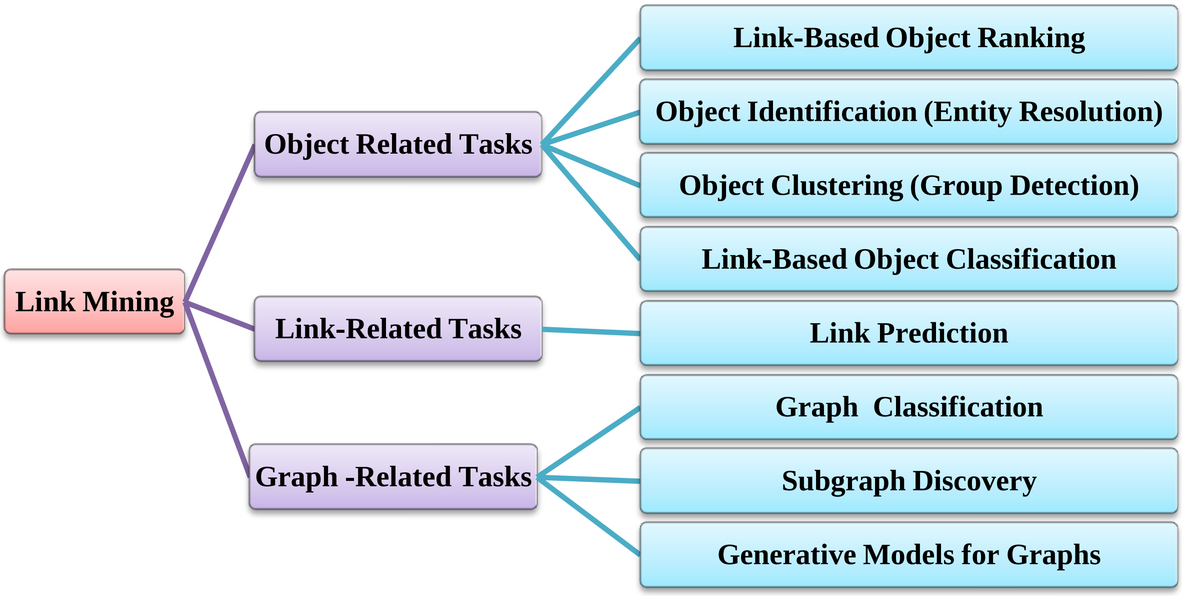
Link Mining
Object Related Tasks
Link-Related Tasks
Graph -Related Tasks
Link-Based Object Ranking
Object Identification (Entity Resolution)
Object Clustering (Group Detection)
Link-Based Object Classification
Link Prediction
Graph Classification
Subgraph Discovery
Generative Models for Graphs

Figure 1.3: Representation of different link mining tasks

1.2.3 Link Prediction in Signed Networks

Link Prediction Problem is a basic computational problem in underlying social network evolution which aims at estimating the likelihood of the existence of a link between two nodes of the network. Now there is a large and rapidly growing literature on the prediction of links in online domains. Some existing approaches are straightforward based on the topology of social networks [Liben-Nowell and Kleinberg, 2007] and showed that information about future interactions can be extracted from the network structure alone. However, few researchers have started to utilize other side of the relationships viz., antagonism along with existing positive relationships. This problem was first studied by [Guha et al., 2004] in a slightly different manner though. In their study, a trust propagation model, exploiting trust and distrust both, was developed with aim to infer trust between unfamiliar users. [Leskovec et al., WWW 2010] first considered an explicit formulation of the sign prediction problem in connection with social psychology theories and uncovered the sought after many effective results that negative relationship can be much helpful even for the tasks involving only positive relationships in the network. A similar standpoint was taken by [Chiang et al., 2011] by exploiting features from longer cycles in social graph in order to benefit the accuracy of sign prediction. Another interesting work was attempted by [Traag and Bruggeman, 2009] to detect the communities in complex networks with the interplay of both positive and negative links.

1.3 Social Recommender Systems

Recommender systems are personalization tools that help users find the right information at the right time based on learnt user preferences. Research on recommender systems has been going on for more than a decade now, but with the increase in the number of e-commerce applications, social networking sites, online users, vendors and increasingly complex products and services, the demand for new intelligent recommendation techniques has also increased dramatically.

“The Web, they say, is leaving the era of search and entering one of discovery. What's the difference? Search is what you do when you're looking for something. Discovery is when something wonderful that you didn't know existed, or didn't know how to ask for, finds you.”

Jeffrey M. O'Brien

The above quote captures the essence of what recommender systems are all about. Recommender systems are personalization tools which enable users to be presented information suiting his interests, which are novel, serendipitous and relevant, without being explicitly asked for. They enable users to present items which they may not know of, thus supporting “discovery” rather than “search”. Recommender Systems have found their way into many entertainment and e-commerce web sites and not only help people to find items of interest but also form communities of interest [Terveen & Hill, 2001]. Recommender systems have become ubiquitous, with their presence everywhere from recommending books (Amazon), CDs, music, movies to even recommending friends.

The increased popularity of social networking applications in recent years has greatly extended active participation and content production of users on SN sites as they often include blogs, photo galleries, videos and other means of sharing digital content. However, they are in danger of becoming victims of their own success with rapidly increasing number of users. Because of the large number of users on the sites identifying like-minded fellow users has become very difficult.

Social Recommender Systems aim to increase adoption, engagement, and participation of new and existing users in social media sites by alleviate information overload over their users by presenting the most attractive and relevant content, often using personalization techniques [Adomavicius and Tuzhilin, 2005], adapted for the specific user. In this work our focus is on a special case of SRS where items to be recommended are friends i.e. Friend recommender system.

1.3.1 Friend Recommender Systems (FRS)

In FRS a user profile is important because it records relevant information such as preferences and behaviour pattern of individual users. User profile stores appropriate approximation of individual's information such as basic information (e.g gender, hometown, language etc) and information of interest which is represented by preferences. List of attributes in profile is necessary step and rest of the analysis depends on it. Once this step is decided, then only other techniques can be defined.

Some factors are most important and strongly affect the relationship in social networks e.g. geographic location is strongest factor affecting how relationships shape up in social network. Presence of attributes such as hometown, city and state in profiles testify to that fact. Studies show that, *"Relationships are more likely to develop between similar individuals."* This theoretical concept is called **Principle of Homophily**, which is related to sociology [McPherson et al., 2001]. Socially connected pairs of individuals are more likely co-resident in the same geographical location, born in the same birth-place and from the same age group.

1.3.2 Recommendations by Utilizing Social Balance Theory

There is a large and rapidly growing literature on the friend recommendation systems in online domains. Few researchers utilized easily accessible social content and compared them with other algorithms to recommend people in an enterprise SN (Chen et al., 2009; Guy et al., 2010). On the other hand, (Agarwal and Bharadwaj, 2011) presented solution of such recommending task using collaborative filtering technique over the various personal and behavioral evolved features and also effectively covered the issues related with low density friend's networks. So far, people have focused only on the use of similarity measures, but (Bian and Holzman, 2011) studied it in a different way through the use of personality matching with collaborative filtering and have established that their approach ensures a higher amount of sustainability in friendship.

However to our knowledge, none of the previous work used the information contained in implicit polarity of the existed links as prior information to suggest a user who are more likely to become his friends/foes in near future. In our work, to carry out investigation with positive and negative links we have used *Social Balance Theory* [Heider, 1946; Cartwright and Harary, 1956; Khanafiah and Situngkir, 2004] that helps us to know about how the structure of SN affects when prediction of new links are made, how the sense of relationships change i.e. a friendly link changes to unfriendly or vice versa and how the balance index serves as an important parameter for the accurate recommendation of new links. By the presence of the balance index we can create feedback over the network balance, so that the system can decide whether to accept or not the arrival of new individual in existing group.

The *social balance theory* is based on the common principles that "the friend of my friend is my friend", "the enemy of my friend is my enemy", "the friend of my enemy is my enemy" and (perhaps less convincingly) "the enemy of my enemy is my friend." Concretely, this means that if w forms a triad with the edge (u, v), then structural balance theory posits that (u, v) should have the sign that causes the triangle on u, v, w to have an odd number of positive signs, just as each of the principles above has an odd number of occurrences of the word "friend."

Our aim in this dissertation work is to estimate affinity parameters of target user towards his friends and foes which in turn enables us to decide his type of relationship with new user. Our model tries to extract information contained in positive as well as in negative links of an individual through inductive learning and utilize social balance theory to avoid possible imbalance in the extended social network.

1.4 Organization of Thesis

This chapter presents a brief overview of the field of social network analysis, recommender systems, link mining and the different tasks related with link mining. **Chapter 2** discusses the link prediction problem; various static measures to solve this problem and presents a framework combining supervised learning and link prediction to classify friend and foes. The proposed link prediction scheme based on inductive learning and social balance theory to recommend friends and foes to target user is presented in **chapter 3**. **Chapter 4** lists the experiments performed and results so obtained. The conclusion and future research directions are discussed in **chapter 5**.

Chapter 2

LINK PREDICTION IN SOCIAL NETWORKS

Link prediction is one of the link mining task is which is very new and fascinating area of research in relational learning field and widely used in social network analysis for various application domains including recommender systems, information retrieval, automatic web hyperlink generation, record linkage, genetic or protein-protein interactions prediction and communication surveillance. Various relational learning methods have been developed to predict the existence of potential links within a relational dataset that typically consists of observed linkages among data objects and attributes of the data objects.

2.1 Problem Description

The link prediction problem is usually described as:

Given a set of data instances $V = \{v_i\}_{i=1}^{n}$ *, which is organized in the form of a social network G = (V, E), where E is the set of observed links, then the task to predict how likely an unobserved link* $e_{ij} \notin E$ *exists between an arbitrary pair of nodes* $\langle v_i, v_j \rangle$ *, in the data network.*

In effect, the link-prediction problem asks: To what extent can the evolution of a social network be modelled using features *intrinsic to the network itself?* The goal is to make this intuitive notion precise and to understand which measures of “proximity” in a network lead to the most accurate link predictions. A number of proximity measures lead to predictions indicating that the network topology does indeed contain latent information from which to infer future interactions.

Link prediction is applicable to a wide variety of areas, such as bibliographic domain, molecular biology, criminal investigations, marketing and recommendation systems. Examples of explicit link prediction problems include automatic Web hyper-link creation [Adafre and Rijke, 2005], genetic or protein-protein interactions prediction and the record linkage problem [Bilenko et al., 2003]. Many well-studied problems can be viewed as a link prediction problem once the data are rendered with a graph/network representation. Such examples are abundant. Information Retrieval can be viewed as dealing with prediction of links between words and documents within a word-document bipartite graph representing word occurrence. Collaborative filtering recommender systems can be viewed as services predicting links between users and items within a user-item bipartite graph representing preferences or purchases [Kautz et al., 1997; Domingos and Richarson, 2001]. Record linkage problem [Winkler, 1994] can be viewed as predicting links among records with same identity and protein/genetic interaction modelling can be viewed as predicting underlying protein/genetic interactions based on interaction networks ob-served from experiments.

Link prediction addresses four different problems as shown in the figure below. Most of the research papers on link prediction concentrate on problem of *link existence.* This is because the link existence problem can be easily extended to the other two problems of *link weight* (links have different weights associated with them) and *link cardinality* (more than one link between same pair of nodes in a social network). The fourth problem of *link type* prediction is a bit different which gives different roles to relationship between two objects [Getoor, 2003].

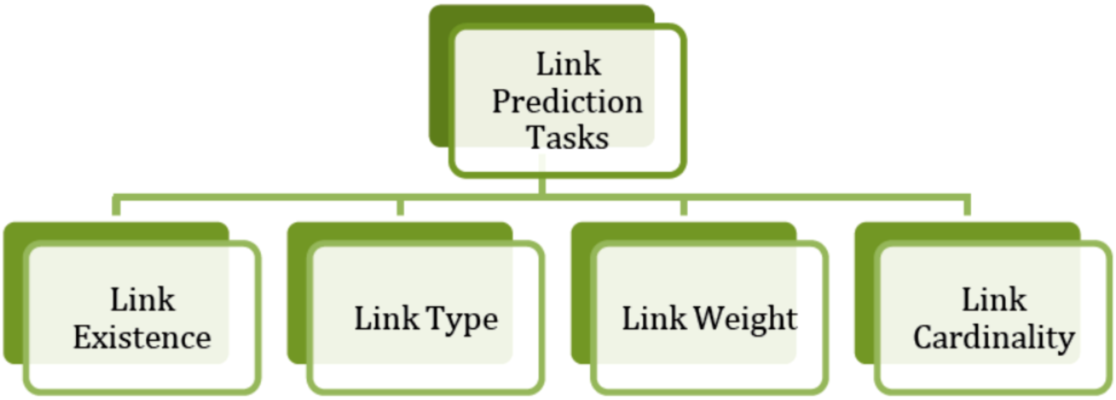

Figure 2.1: Differentiation of link prediction tasks

2.2 Link Prediction Techniques

There is a variety of techniques for the link prediction problem, shown below in fig. 2.2, ranging from graph theory, metric learning, statistical relational learning, probabilistic graphical models and classic classification methods like K nearest neighbors, SVM, multilayer perceptron [Hasan et al, 2006] etc. The link prediction models can mainly fall into four categories in accordance with the intuitions of their solutions: First the node-wise similarity based methods, which focus on seeking a similarity measurement to determine the link existence. Second the topological pattern based methods which try to exploit topological pattern, ranging from local patterns around the nodes to the global patterns covering the entire social network. Third is to use probabilistic model based approach where the idea is to learn a model composed of a set of parameters θ, given the observed social network, according to some optimization strategies and find linked node pairs with this model. And the last is supervise learning approaches where most of the well-known classification algorithms (decision tree, k-nn, multilayer perceptron, SVM, rbf network) can predict link with surpassing performances.

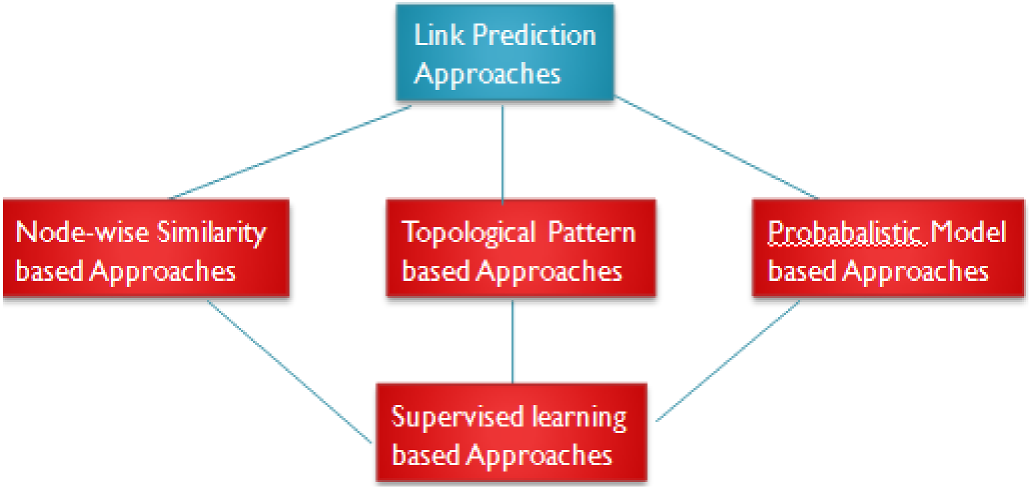
Link Prediction Approaches
Node-wise Similarity based Approaches
Topological Pattern based Approaches
Probabalistic Model based Approaches
Supervised learning based Approaches

Figure 2.2: Different approaches to Link Prediction

2.2.1 Node Wise Similarity Based Approaches

Given an arbitrary pair of data instances (v, u) from the social network G, where the content of each instance is represented by a feature vector Xα = (xα1, xα2, ..., xαn), the node-wise similarity based approaches set their targets as seeking some similarity measurement Sim(***X***α, ***X***β) for the pair of vertices, and then link prediction is achieved by putting an edge between vertices that are at a similarity from each other larger than a fixed threshold δ.

Demographic filtering is very good example for illustrating this method where attributes of nodes is used to find similarity between them. In friend recommender systems a list of attributes like Hometown, Religion, Age, Language, Cuisines, Passions, Activities is used to measure similarity between two users. The popular formula to measure the similarity in recommender system is Pearson Correlation Coefficient (PCC).

Pearson correlation coefficient, $$sim(x,y)=\frac{\sum_{s\in S_{xy}}\left(r_{x,s}-\bar{r}_x\right)\left(r_{y,s}-\bar{r}_y\right)}{\sqrt{\sum_{s\in S_{xy}}\left(r_{x,s}-\bar{r}_x\right)^2\sum_{s\in S_{xy}}\left(r_{y,s}-\bar{r}_y\right)^2}}$$

S_{xy} : set of all items corated by both users *x* and *y*

$r_{u,i}$ = rating of user u on item *i*

2.2.2 Topological Pattern based Approaches

Given an arbitrary pair of vertices (x, y) in a social network G = (V, E), the topological pattern based approaches try to exploit some topological patterns, either local or global, from the observed part of the network. Then the decision for link existence is made by calculating a connection weight score(x, y) for the pair of nodes (x, y) based on the discovered topological patterns. The different techniques are shown in Fig. 2.2

2.2.2.1 Node based Topological Patterns

The node based topological patterns only take the information around each node into consideration. For a node x, let Γ(x) denote the set of immediate neighbors of x in G.

- ***Common neighbors***. The most direct implementation of this idea for link prediction is to define $\mathbf{score}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\Gamma}(\boldsymbol{x}) \cap \boldsymbol{\Gamma}(\boldsymbol{y})$, the number of neighbors that *x* and *y* have in common.
- ***Jaccard's coefficient***. The Jaccard coefficient, a commonly used similarity metric in information retrieval measures the probability that both x and y have a feature f, for a randomly selected feature f that either x or y has. If we take features here to be neighbors in G, this approach leads to the measure:

$$\mathbf{score}(\mathbf{x}, \mathbf{y}) = \frac{\boldsymbol{\Gamma}(\mathbf{x}) \cap \boldsymbol{\Gamma}(\mathbf{y})}{\boldsymbol{\Gamma}(\mathbf{x}) \cup \boldsymbol{\Gamma}(\mathbf{y})}$$

- ***Adamic/Adar*** . [Adamic and Adar, 2003] considered a similar measure, which refines the simply counting of common features by weighting rarer features more heavily. This idea suggests the measure as:

$$\mathbf{score}(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{z} \in \boldsymbol{\Gamma}(\mathbf{x}) \cap \boldsymbol{\Gamma}(\mathbf{y})} \frac{\mathbf{1}}{\mathbf{log}\ |\boldsymbol{\Gamma}(\mathbf{z})|}$$

- ***Preferential attachment.*** The basic premise is that the probability that the probability of co-authorship of x and y is correlated with the product of the number of collaborator of x and y. This proposal corresponds to the measure:

$$\mathbf{score(x, y) = |\Gamma(x)| \cdot |\Gamma(y)|.}$$

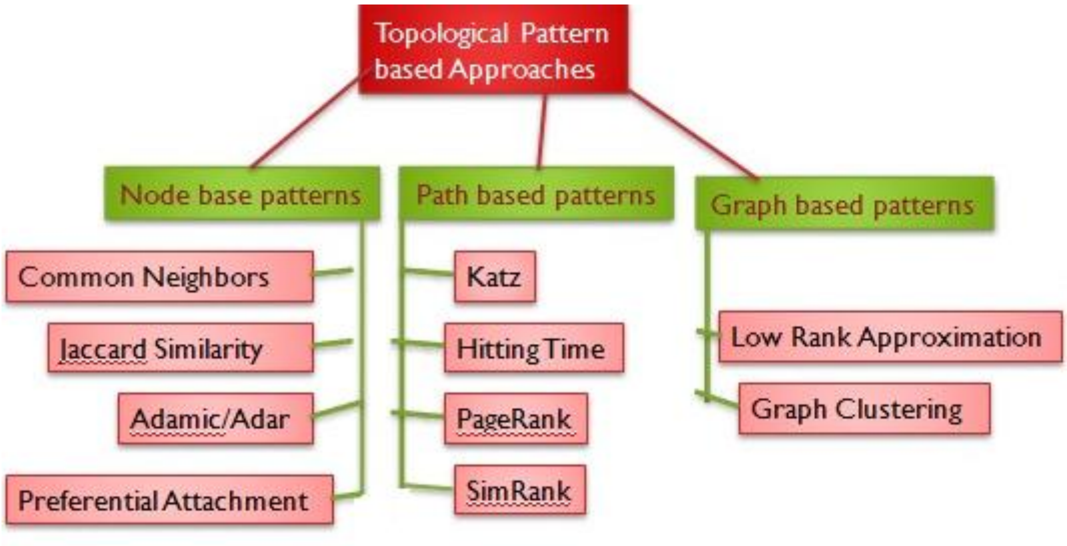

Figure 2.3: Topological pattern based Link Prediction approach.

2.2.2.2 Path based Topological Patterns

The path based topological patterns describe the connectivity between any pair of nodes by the ensemble of all paths between them. Thus, they involve a wider range of information contained in the social network. The key idea shared among these approaches is that, if there are many paths indirectly connecting node x with node y, it is likely that there is a link connecting them directly. In this section, we list a number of methods which refine the notion of shortest-path distance by implicitly considering the ensemble of all paths between two nodes.

- ***Katz.*** Katz defined a measure that directly sums over this collection of paths, exponentially damped by length to count short paths more heavily. This notion leads the measure where $\mathbf{paths}_{\mathbf{x,y}}^{\langle l \rangle}$ is the set of all length-l paths from x to y, and β > 0 is parameter of the predictor.

$$\mathbf{score}(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{l=1}}^{\infty} \boldsymbol{\beta}^{\mathbf{l}} |\mathbf{paths}_{\mathbf{x,y}}^{\langle l \rangle}|$$

- ***Hitting time***. A random walk on G starts at a node x and iteratively moves to a neighbor of x chosen uniformly at random from the set Γ(x). The hitting time $H_{x,y}$ from x to y is the expected number of steps required for a random walk starting at x to reach y. Because the hitting time is not in general symmetric, it also is natural to consider the commute time $\mathbf{C_{x,y} = H_{x,y} + H_{y,x}}$. Both of these measures serve as natural proximity measures and hence (negated) can be used as score(x, y).

One difficulty with hitting time as a measure of proximity is that Hx,y is quite small whenever *y* is a node with a large *stationary probability πy*, regardless of the identity of *x*. To counterbalance this phenomenon, we also consider *normalized* versions of the hitting and commute times, by defining

$$\mathbf{Score}(\boldsymbol{x},\boldsymbol{y}) = -(\boldsymbol{H_{x,y}} \,.\, \boldsymbol{\pi_{y}} + \boldsymbol{H_{y,x}} \,.\, \boldsymbol{\pi_{x}}).$$

- ***PageRank.*** It can be used for link prediction. Define score(x, y) under the rooted PageRank measure with parameter α ∈ [0, 1] to be the stationary probability of y in a random walk that returns to x with probability α each step, moving to a random neighbor with probability 1 − α.

- ***SimRank***. SimRank is a fixed point of the following recursive definition: Two nodes are similar to the extent that they are joined to similar neighbors. Numerically, this quantity is specified by defining score(x, x) = 1 and

$$\mathbf{score}(\mathbf{x},\mathbf{y}) = \boldsymbol{\gamma}.\frac{\sum_{\mathbf{a}\in\boldsymbol{\Gamma}(\mathbf{x})}\sum_{\mathbf{b}\in\boldsymbol{\Gamma}(\mathbf{y})}\mathbf{score}(\mathbf{a},\mathbf{b})}{\boldsymbol{\Gamma}(\mathbf{x}).\boldsymbol{\Gamma}(\mathbf{y})}$$

  for a parameter γ ∈ [0, 1].

2.2.2.3 Graph based Topological Patterns

Graph factorization based methods try to approximating the adjacency matrix M of the data graph G by the product Mk of some low-rank matrices, which reveal some global structural patterns of the observed social network.

- ***Low-rank approximation:*** A common general technique when analyzing the structure of a large matrix *M* is to choose a relatively small number *k* and compute the rank-*k* matrix M_k that best approximates *M* with respect to any of a number of standard matrix norms. This computation can be done efficiently using the singular-value decomposition. [Liben-Nowell and Kleinberg, 2007] investigated three applications of low-rank approximation: (i) ranking by the Katz measure, in which M_k is used rather than M in the underlying formula; (ii) ranking by common neighbours, in which score is calculated by inner products of rows in M_k rather than M; and most simply of all (iii) defining score(x, y) to be the ‹x, y› entry in the matrix M_k.

- ***Clustering***. One might seek to improve on the quality of a predictor by deleting the more “tenuous” edges in graph *G* through a clustering procedure, and then running the predictor on the resulting “cleaned-up” subgraph. Consider a measure computing values for score(*x*, *y*). Compute score(*u*, *v*) for all edges in G and delete the (1-ρ) fraction of these edges for which the score is lowest, for a parameter ρ ∊ [0, 1] . Now recomputed score(*x*, *y*) for all pairs *x*, *y* on this subgraph; in this way, determine node proximities using only edges for which the proximity measure itself has the most confidence.

2.2.3 Probabilistic Model Based Approaches

Relational modelling has recently received increasing attention and plays an important role in modern data mining. The reason is that it could encapsulate relevant information contained either in single objects, relationships or the underlying structure of the entire data network. With the learned model, both the vertices and the edges in the data graph can be re-generated. Two of the leading frameworks, i.e. the *Probabilistic Relational Models (PRM)* framework [Friedman et al., 1999] and the *Directed Acyclic Probabilistic Entity Relationship (DAPER)* framework [Heckerman et al., 2004], describe relational modelling in the context of relational databases. They are motivated from different database structure representations: the PRM model is based on the *relational model* and the DAPER model is based on the *entity-relationship model*.

2.2.4 Supervised Machine Learning Approach

In this approach, the goal is to discriminate between examples of the linked class (positive examples) against examples of the not-linked class (negative examples). Learning such a supervised classification model requires building a training data that describes examples of both classes.

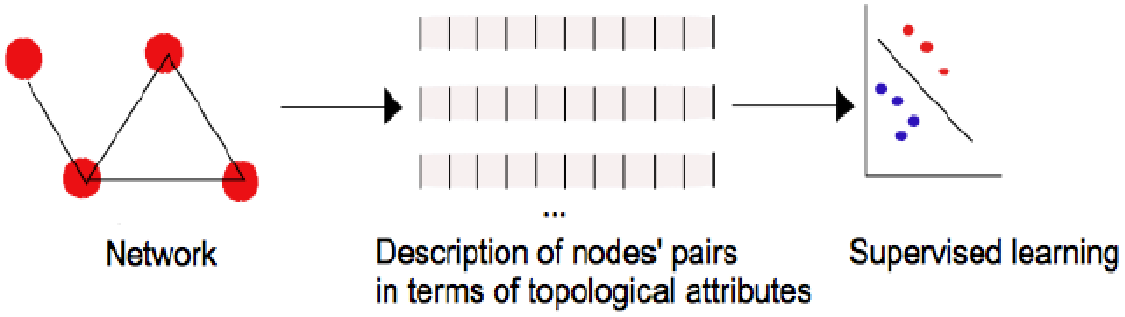
Network
...
Description of nodes' pairs in terms of topological attributes
Supervised learning

Figure 2.3: Reformulating Link Prediction as a supervised learning task

To use supervised learning approach for link prediction we have to follow the steps as given below:

- Training data Preparation
- Feature Set Selection
- Classification Algorithm Selection

2.3 Framework for Link Prediction Using Decision Tree Learning

There exists a plethora of classification algorithms for supervised learning. Although their performances are comparable, some usually work better than others for a specific dataset or domain. We employ decision tree learning to design the classifier as it is best suited to problem used in my work.

As mentioned above link prediction is reformulated simply as classification problem if we want to use supervised approach. A classification technique (or classifier) is a systematic approach which is used to build classification models by adopting a learning algorithm to accurately predict the class labels of previously unknown records. In this work, C4.5, a decision tree based classification algorithm, is employed for the desired classification [Mazid et al., 2010].

A decision tree classifies instances by traversing down the tree from the root to some leaf node which denotes a particular classification. All leaf nodes in a decision tree provide for a classification and all the non-leaf nodes in a decision tree represent a particular attribute of the instance. The branches descending from an attribute node denote the values taken by the particular attribute. There can be more than one decision tree that is consistent with the given data, whereby arises the question of choosing the most appropriate tree. Our decision tree learning algorithm relies on the Ockham's razor in tackling this issue. According to the Ockham's razor: "Prefer the simplest hypothesis consistent with the data", which can be understood in decision tree learning terms as selecting the shortest possible decision tree. Our decision tree learning algorithm is based on the ID3 algorithm [Quinlan, 1986].

2.3.1 Modeling of User's Profile Information

Formally, we represent our dataset as undirected graph = (V, E) where V is a finite set of vertices or nodes of the graph and E consists of set of edges of the form (*i, j*) such that *i, j*∈ V, with a signed (positive or negative) notations. All the signed entries are collectively stored in a n×n user-user matrix denoted by A, where each positive and negative entry is replaced by +1 and -1 respectively. In the special case where no edge exists between *i* and *j*, it would be given value 0. We label the nodes to which active user creates positive links as friends $f_r = (f_{r1}, f_{r2}, \dots, f_{rm})$ while to other nodes u creates negative links as foes $f_o = (f_{o1}, f_{o2}, \dots, f_{on})$. Moreover, L={ $l_1, l_2, \dots, l_k$ } is the set of nodes to which active user *u* is not yet connected. First we begin by defining an appropriate set of attributes and hence attribute values and categorical information, required to induce decision tree and rules. To demonstrate our scheme, we have chosen a list of few attributes/Category as given below:

Attributes: *Gender, Career Interest, Hometown, Movies, Thinking,Religion, SES, Activities.*
Categories: *Friend &Foes*

Next, we generate a profile for every user, where each tuple corresponds to a link existing between him and others on social networks, using above defined attributes along with the class label (friend or foe).

2.3.2 Inductive Decision Tree Based Learning

Now, our task is to build a decision tree from the input vectors using the above attributes-categories structure and we generate classification rules by employing C4.5. Applicability of the rules, generated from this algorithm, over the information of unknown user can easily make its classification to either of the classes.

C4.5, a tool to draw decision tree based on ID3 algorithm, constructs a very big tree by considering heuristic information gain approach, includes all attribute values and finalizes the decision rule by pruning. An algorithm for construction of C4.5 decision tree is as follows:

- *Select the attribute associated with each node of decision tree which is most informative among the attributes not yet considered in the path from the root.*

- *Then pass away collected information to subsequent nodes, called 'branch nodes' which eventually terminate in leaf nodes that will give decisions.*
- *Repeat the above steps until all samples are exhausted or no attribute values are left or all samples belongs to same class.*
- *Next, the process to remove least reliable branches which generally results in faster classification using some statistical measure is done.*
- *Finally, extracts rules (IF-THEN rules) from decision tree by illustrating the path from the root node to leaf node.*

In this work, we are modeling our link prediction problem as classification task of unknown links (i.e. are they positive or negative links) based on rules generated by considering the values of attributes associated with profiles of users in friends and foes networks (FFN) using the above described inductive learning decision tree based classification approach. Thereafter, we recommend positive and negative links, using the extracted rules, to an individual avoiding possible social imbalance in the extended friends & foes network of an individual based on social balance theory.

Chapter 3

PROPOSED LINK PREDICTION SCHEME

Social networks are the patterns of contact that are created by the flow of messages among communicators through time and space. The concept of message should be understood here in its broadest sense to refer to data, information, knowledge, images, symbols, diseases, infections, signals and any other symbolic forms that can move from one point in a network to another or can be co-created by network members. These networks take many forms in contemporary organizations, including personal contact networks, flows of information within and between groups, strategic alliances among firms, and global network organizations, to name but a few [Monge and contractor, 2003]. In real life these networks are formed because of some motives, intentions or say reasons. Social science researchers collectively call those motives as *social science theories*. Social networks are highly dynamic in nature and they evolve and changes continuously to fulfil the need of individuals. To bring virtual online social networks more closer to real social networks the computer scientist are trying to incorporate these theories in making the computation models for simulating the OSN. The fig 3.1 below is depicting the same idea.

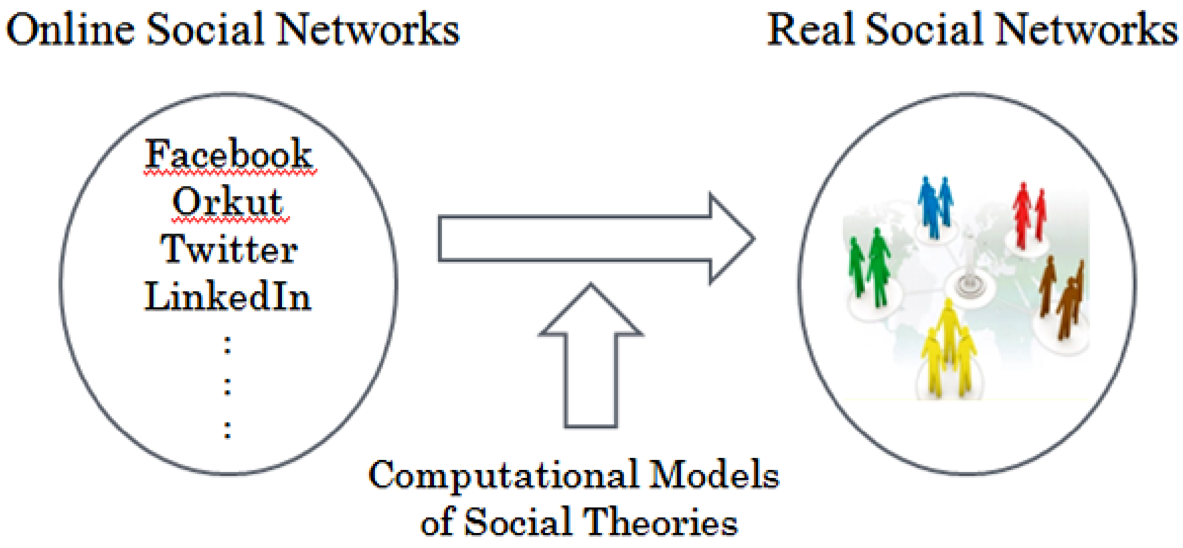

Figure 3.1: Incorporation of social theories in OSN to simulate real social network

Some of these social theories are:

- *Social balance theory*
- *Theories of self-interest*
- *Contagion theory*
- *Social influence theory*
- *Exchange and dependency theory*
- *Homophily and proximity theory*

In recent times many authors from computer science field have used this theory in their papers. Leskovec et al wrote two of the paper using the social balance theory [Leskovec et al, CHI 2010; Leskovec et al, WWW 2010]. [Ahmad et al, 2010] used insights from these theories of evolution of social communication networks and the Multi-Theoretical Multi-Level (MTML) framework, which synthesized insights from the theories mentioned above, to derive models which can be used to make link predictions across networks.

We are using Social balance theory in our dissertation work. For applying this theory it is mandatory to have a signed network so that we can balance that network to keep relations stable and smooth. There is by now a large and rapidly growing literature on the analysis of social networks arising in on-line domains, this line of work has almost exclusively treated networks as implicitly having positive relationships only. But many times social interaction involves negative relationships besides the positive one especially in social media sites.

3.1 Positive And Negative Links in Social Networks

Positive links indicate friendship, support, or approval and negative link signify disapproval of others, or express disagreement or distrust of the opinions of others. For instance, in on-line rating sites such as Epinions, people can give both positive and negative ratings not only to items but also to other raters. In on-line discussion sites such as Slashdot, users can tag other users as “friends” and “foes”. In Wikipedia one user can vote for or against on the promotion to admin status of another.

Till now, most of the work has been done using positive links except a few. Leskovec et al. worked on edge sign prediction problem which is closely related with link prediction. He studied the dataset from Epinion, Slashdot and Wikipedia and found that sign of links in underlying social network can be predicted with high accuracy using models that generalize across this diverse range of sites. [Kunegies et al., 2010] analyse the corpus of user relationship of the Slashdot technology news site and identified unpopular users.

To carry out an investigation with positive and negative links we need theories of signed networks that help us reason about how different patterns of positive and negative links provide evidence for the expression of different kinds of relationships. *Social balance theory* is one of such theory.

3.2 Social Balance Theory

This theory is formulated by Heider in 1946 and subsequently cast in graph-theoretic language by Cartwright and Harary in 1956 [Khanafiah and Situngkir, 2004]. Structure balance considers a simple social network with three individuals (i. e., a triad) in which each pair of actors is connected by either a positive or negative link. This links represent the sentiments of the actors toward each other [Antal et al., 2006]. Ignoring the identities of the actors, there are four possible types of triads, shown below. Here Solid lines denote positive links, and dotted lines denote negative links and the index for each triad simply indicates the number of negative links.

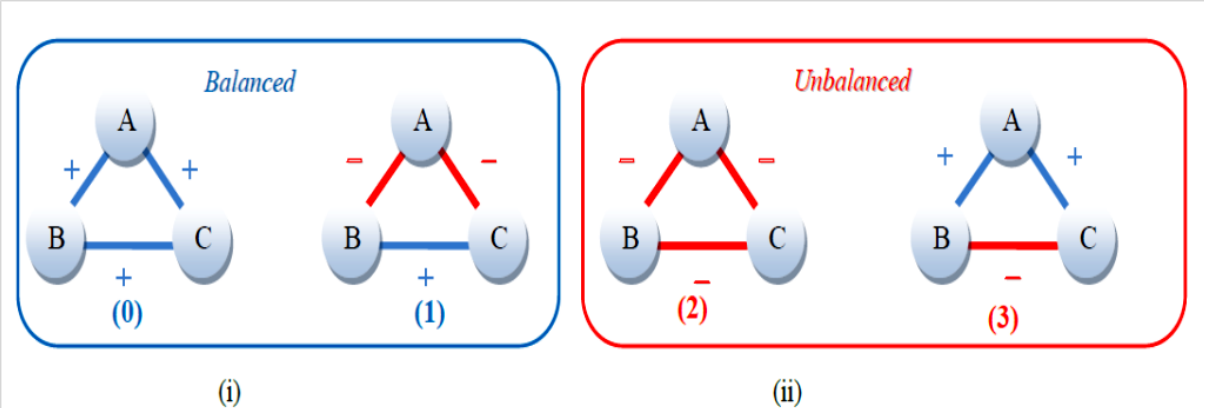

Figure 3.2**:** Undirected signed triads

In the triad, balance state occurs when all sign multiplication of its sentiment relation charges positive. In this way, balance state will occur when there are sentiment relations with signs all positive (+ x + x + = +), or two negatives and one positive (- x – x + = +). So triad 0 and triad 2 are balanced and triad 1 and triad 3 are unbalanced. The balanced triad 0 can be interpreted as *my friend's friend is my friend.* Similarly balanced triad 2 can have any one of the three interpretations- *my friend's enemy is my enemy*, *my enemy's friend is my enemy* or *my* *enemy's enemy is my friend.* Essentially, type-1 triads are imbalanced because they violate the principle that "*a friend of my friend is my friend*" and type-3 triads are imbalanced because they violate the principle that "*an enemy of my enemy is my friend*"

A complete signed graph is balanced if and only if every triad within this graph is balanced. In Fig. 3.2 the one on the left is balanced, because each triad is balanced. On the other hand, the one on the right is not balanced, because triad ABC and BCD are unbalanced**.**

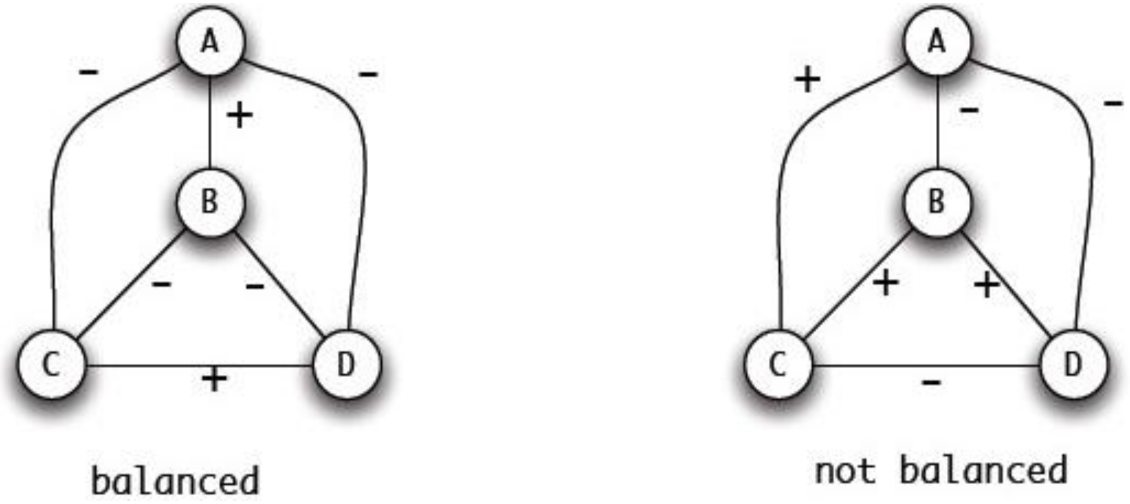

Figure 3.3: Example of balanced and unbalanced signed graph

Balance can also be assessed in a second way. This method is popularly known as Structure theorem and is given by Cartwright and Harary (1956).

Structure Theorem

A graph (network of individuals) is balanced iff the group can be divided into two subgroups (two sets), wherein individual relations in the same subgroup are all positive (all edges between vertices in the same set are '+') and between individuals in different subgroups are negative.

Intuitively, the network is balanced when the actors can be separated into (no more than) two "camps" in such a way that any friends belong to the same camp and any enemies belong to different camps. Returning to our 4 types of triads, note that this partition is possible for triads of type 0 (where all actors belong to the same camp) and type 2 (where 2 actors belong to one camp and the remaining actor belongs to the other camp). However, this partition is not possible for triads of type 1 or 3.

3.2.1 The Model

Every interpersonal network tends towards higher balance. A simple measure of the “degree” of balance is given by the proportion of triads that are balanced. So the global balance index, β can be written as:

$$\beta = \frac{T_{balanced}}{T_{total}} \qquad (1)$$

where $T_{balanced}$ denotes the number of balanced triads, T_{total} denotes the total number of triads in the whole interpersonal network.

By the presence of the balance index we can create feedback over the network balance, so that the system can decide whether accept or not the arrival of new individual in existing group. The assumption that a network tends toward higher balance brings the mechanism of acceptance or rejection of the change. The new individual will be accepted if the balance index is higher than the previous one.

In a group consists of *N* individuals, the number of dyads (possible sentiment relations), denoted by *D*, equals to

$$D = \frac{N!}{(N-2)!\,2!} \qquad (2)$$

If the possible types of dyadic relations formed are *3* (whether positive, negative, or no-relation), then the possible relation patterns (*p*) are:

$$\mathrm{p} = 3^{\mathrm{D}} = 3^{(\frac{\mathrm{N!}}{\mathrm{(N-2)!2!}}} \qquad (3)$$

Whereas the number of individuals combination which formed triads in the group consists of N individuals are:

$$T_{total} = \frac{N!}{(N-3)!\,3!} \qquad (4)$$

An adjacency matrix of *N x N*, is used to draw pattern of relation (edges) formed between connected individuals (vertices), where *R_{ij}* is interconnectivity sign of *i*-individuals over *j*-individuals and vice versa, that can be positive, negative, or no-relation, signed with *+1* for positive relation, *-1* for negative relation, and *0* for no-relation. From the matrix, the possible triads can be determined to be balance or not, and we can also determine the balance index of the network.

3.2.2 Balance Index (BI) computation

To compute balance index of a network (or graph) we need to count total number of balanced and unbalanced triads. So our basic problem here is the computation of triangles. There may be various literature giving procedures and formulae for the triangle computation. Here I am describing one of the methods to do the same.

Representation of Network

We represent the network as Adjacency matrix. As we are using signed network so there will be three kind of entries, 1 for positive link, -1 for negative link and 0 if there is no link between the two nodes. Also we are using undirected graph so the adjacency matrix will be symmetric. To avoid redundancy and make computation easy we store only upper triangular matrix. Let us denote this adjacency matrix as G. Now divide this matrix into two separate matrices P and N such that P + N = G; P is having only positive links and N having only negative links.

Now for our computation we have to calculate all four kinds of triads and these are:

(1) + + + (balanced)

(2) + + -, + - +, or - + + (unbalanced)

(3) + - -, - + -, or - - + (balanced)

(4) - - - (unbalanced)

Algorithm for computing Balance Index (BI)

```
S1. PP= P*P (each entry will give number of paths between a pair of nodes of length two,++)
S2. PN= P*N (each entry will give number of paths between a pair of nodes of length two,+-)
S3. NP= N*P (each entry will give number of paths between a pair of nodes of length two,-+)
S4. NN=N*N (each entry will give number of paths between a pair of nodes of length two, --)
//computing number of triads in (1)
S5.  PPP= P.* PP;   (triads + + +)
S6. S1=sum(PPP(:)); (S1 is the number of triads of type +++ in graph. )
//computing number of triads in (2)
S7. PPN= P.*(PN) (triads ++-)
S8. PNP= P.*(NP) (triads +-+)
S9. NPP=N*.(PP)  (triads  - ++).
S10.  T2= PPN + PNP + NPP
S11.   S2= sum(T2(:))
//computing number of triads in (3)
S12. NPN= N.*(PN) (triads -+-)
S13. NNP= N.*(NN) (triads --+)
S14. PNN=P*.(NN)  (triads + - -).
S15.  T3= NPN + NNP + PNN
S16.   S3= sum(T3(:))
//computing number of triads in (4)
S17.  NNN= N.* NN;   (triads ---)
S18. S4=sum(NNN(:));
```

Balance index (BI) = $\beta = \frac{T_{balanced}}{T_{total}} = \frac{S1+S3}{S1+S2+S3+S4}$

3.3Main Steps of the Proposed Framework

***Step1*:** Applying an inductive learning classification approach to build a classifier that utilizes information contained in $f_r = (f_{r1}, f_{r2}, \dots, f_{rm})$ and $f_o = (f_{o1}, f_{o2}, \dots, f_{on})$ sets of active user.

Step 2: Then classifier identifies set of possible users L= $\{l_1, l_2, \dots, l_k\}$ who can be friend/foes of active user.

Step 3**:** Next, computes balance index (BI) for every user presented in set L using the above mentioned Formula (1).

Step 4**:** Finally, recommendations of set of users which consists of friends/foes are made which either maintains or enhances the balance index of the FFN of the active user.

The scheme which facilitates the understanding of generated recommendation of Friends/foes for an individual using balance theory is presented in Fig. 3.4.

Algorithm:

```
INPUT: current FFN of active user, set of possible N new links obtained through the classifier.
OUTPUT: Set of $k$ recommended FFLs
(1) Compute BI of active user's current FFN i.e. BI(current FFN)
(2) Initialize $\beta \leftarrow$ BI (current FFN)
(3) Repeat
   a) Extended FFN = FFN after adding k=N new links to the current FFN
   b) Compute BI(extended FFN) i.e. the balance index of active user's extended FFN
   c) If (BI (extended FFN)$\geq$ BI (current FFN))
            $\beta \leftarrow$ BI (extended FFN)
            exit;
        else
           $k = N-1$;
        Until $k \geq 1$
```

Fig 3.4 Process for generating recommendations of new FFLs set using social balance theory.

Chapter 4

EXPERIMENTAL RESULTS AND ANALYSIS

In this chapter we present the results of conducting experiments using the methods proposed in this work. The aim is to recommend Friends and Foes to user after filtering with social balance theory. We will recommend only those connections which ultimately retain or increase the balance index of network of the target from its current balance index and at same time was the maximal set. The experiments that we have conducted uses simulated dataset of 25 users.

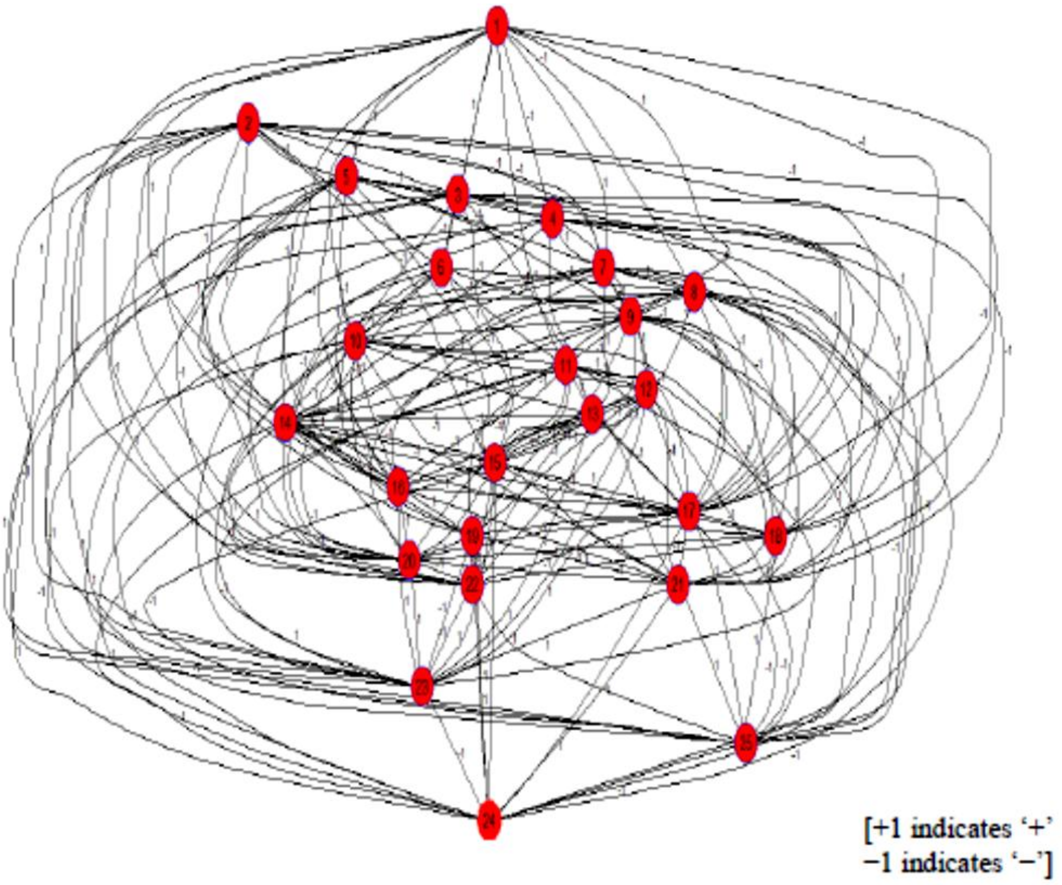

Figure 4.1: A signed social network of 25 users

5.1 The Dataset

To train our classifier model the dataset which we have used consists of set of profiles corresponding to every individual. Each profile contains set of records, wherein a single record has a set of attribute-value pairs along with the class of relationship. Most of the attributes exhibited by user in the dataset are discrete valued attributes. The data which we have collected represents a small social network (Fig. 4.1), with ties existing among people only who are in relationships. Also, each tie is annotated with either a positive or a negative sign depending on the type of relation. A sample profile of an active user is given in Table 4.1. Here, we strictly assumed that each user must be connected to at least 55% of nodes in the network. We learn a classifier for every individual and then apply classifier over the remaining 45% of dataset. One important property of our scheme is that its scope in finding new FFLs is not only limited to 2-hops social neighbourhood i.e. friends of friends, but it has global presence.

Table 2.1: A sample profile of an active user 1 with attribute-value pairs and class label information

User	Gender	Hometown	Career Interest	movies	activities	religion	Thinking	SES	like
1(Active user)	**F**	**delhi**	**politics**	**sci-fi**	**cooking**	**hindu**	**moderate**	**high**	**friend**
2	M	bihar	academic	sci-fi	sports	sikhh	liberal	high	Unknown
3	F	punjab	R&D	historical	drawing	sikhh	liberal	high	friend
4	F	delhi	R&D	comedy	chatting	bodhh	moderate	low	foe
5	M	up	corporate	action	singing	sikhh	orthodox	medium	friend
6	M	punjab	corporate	romance	writing	bodhh	liberal	high	friend
7	F	w.bengal	politics	action	drawing	hindu	moderate	medium	friend
8	F	punjab	politics	romance	sports	sikhh	orthodox	low	foe
9	M	mp	corporate	comedy	singing	bodhh	orthodox	medium	friend
10	F	bihar	corporate	action	reading	christian	orthodox	high	foe
11	M	delhi	R&D	historical	surfing	hindu	moderate	medium	friend
12	M	mp	corporate	comedy	surfing	bodhh	liberal	high	Unknown
13	F	mp	R&D	comedy	singing	christian	liberal	low	Unknown
14	M	delhi	politics	sci-fi	cooking	hindu	liberal	medium	foe
15	F	bihar	corporate	sci-fi	reading	hindu	orthodox	low	Unknown
16	F	up	admin	horror	sports	muslim	moderate	high	friend
17	F	w.bengal	academic	historical	drawing	muslim	orthodox	medium	foe
18	M	rajsthan	corporate	comedy	shopping	christian	moderate	high	Unknown
19	M	haryana	corporate	horror	cooking	christian	orthodox	low	Unknown
20	M	mp	corporate	sci-fi	surfing	sikhh	moderate	low	Unknown
21	M	up	corporate	romance	singing	hindu	moderate	medium	Unknown
22	M	up	corporate	horror	chatting	hindu	moderate	high	foe
23	F	rajsthan	admin	action	drinking	muslim	orthodox	medium	foe
24	F	punjab	politics	romance	reading	sikhh	orthodox	high	Unknown
25	M	delhi	admin	action	surfing	muslim	liberal	medium	foe

4.2 Experimental Results and Discussion

To illustrate the details of our proposed scheme, initially we begin with a particular user, specified as active user (User 1) which has 7 friends and 8 foes (Table 4.1). Out of 25 users, 15 were chosen as training users, user 1 is the active user and remaining 9 are unknown users for which prediction is to be made. After collection of their attributes and link information, our next focus is to learn a classifier for active user which is further used to predict the sign of unknown links.

Following are the discovered rules corresponding to the active user 1 based on data set in Table 4.1:

```
Rules:

  Rule 1: (2, lift 1.6)
          hometown = delhi
          thinking = liberal
          ->  class foe  [0.750]

  Rule 2: (10/2, lift 1.6)
          gender = F
          thinking = orthodox
          ->  class foe  [0.750]

  Rule 3: (6/1, lift 1.4)
          hometown = punjab
          ->  class friend [0.750]

  Rule 4: (12/3, lift 1.3)
          thinking = moderate
          ->  class friend [0.714]

  Rule 5: (12/4, lift 1.2)
          gender = M
          thinking = liberal
          ->  class friend [0.643]
```

In the graph (Fig. 4.1) 2, 12, 13, 15, 18, 19, 20, 21 and 24 are the nodes to which the active user 1 is not connected yet. The above set of rules, classifies only 6 out of these 9 unknown users i.e. {15, 18, 19, 20, 21, 24}, and categorized as foe, friend, friend, friend, friend and foe respectively. Therefore, the possible set of new FFLs is {1−>15(-), 1−> 18(+), 1−>19(+), 1−>20(+), 1−>21(+), and 1−> 24(−)}. Next, we iteratively process all the subset of above set of FFLs to choose effective sets of new links that either *maintains* or *enhances* the balancing index of the extended FFN of the active user. Table 4.2 shows different order sets obtained when social balance theory aspects are applied over the active user network.

From Table 4.2, it is clear that the final set of links to be recommended by the system would be {15, 18, 19, 20, 21, 24} which satisfies both the criteria i) It is the largest possible set, and ii) connection of such new links in the current network increases the balance index of the extended FFN of active user.

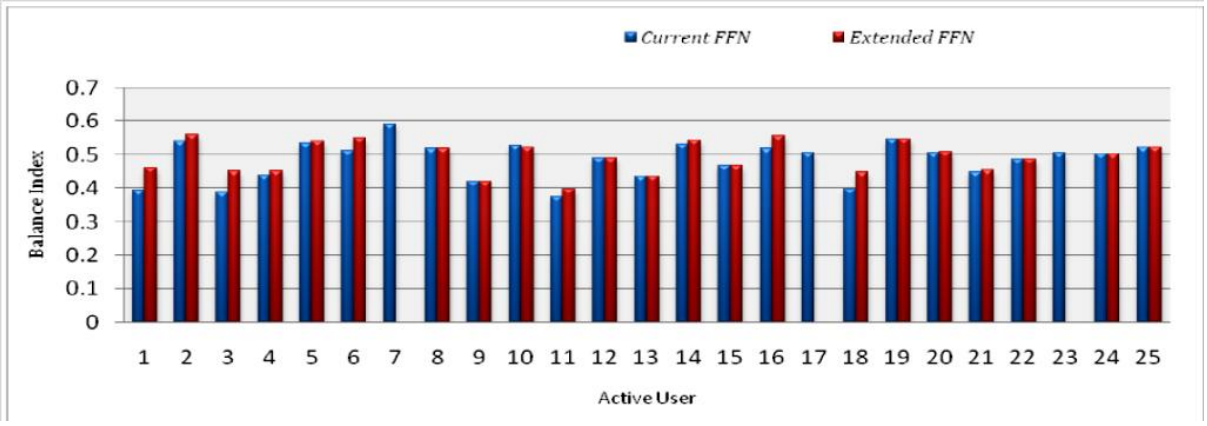

Figure 4.2: Comparison of BI of current and extended FFN of active user

Table 4.2: Representation of sets of unknown links for active user which satisfies the recommendation criteria of balance theory

Set order	***BI(Extended FFN)> BI(Current FFN)***	***BI(Extended FFN) = =BI(Current FFN)***
{First order}	(15), (18), (19), (20), (24)	
{Second order}	(15,18), (15,19), (15,20), (15,21), (15,24), (18,19), (18,20), (18,24), (19,20), (19,21), (19,24), (20,21), (20,24)	
{Third order}	(15,18,19), (15,18,20), (15,18,21), (15,18,24), (15,19,20), (15,19,21), (15,19,24), (15,20,21), (15,20,24), (15,21,24), (18,19,20), (18,19,21), (18,19,24), (18,20,21), (18,20,24), (18,21,24), (19,20,21), (19,20,24), (19,21,24), (20,21,24)	(18,21,24)
{Fourth order}	(15,18,19,20), (15,18,19,21), (15,18,19,24), (15,18,20,21), (15,18,20,24), (15,18,21,24), (15,19,20,21), (15,19,20,24), (15,19,21,24), (15,20,21,24), (18,19,20,21), (18,19,20,24), (18,19,21,24), (18,20,21,24), (19,20,21,24)	
{Fifth order}	(15,18,19,20,21), (15,18,19,20,24), (15,18,19,21,24), (15,18,20,21,24), (15,19,20,21,24), (18,19,20,21,24)	
{Sixth order}	(15,18,19,20,21,24)	

0.3939− BI of current FFN of active user 1

In the next part of our analysis, we perform the same process, as previously described for a single user, over the entire network. The different order sets and their corresponding balance index that we have obtained are graphically presented in Fig. 4.2. There are some cases where no such sets are found that can fulfill the recommendation criteria of balance theory, although they were classified successfully by the discovered set of rules. In Fig. 4.2, user 7, 17 and 23 indicate the situation of non availability of these sets. The final recommended sets of FFLs corresponding to every individual in this network are listed in Table 4.3.

Table 4.3: List of Recommended FFLs

user	**Recommended set**
1	{15,18,19,20,21,24}
2	{1,4,8,11,12,13,21}
3	{7,8,9,12,23}
4	{6,8,10,12,13,24}
5	{8,12,14 }
6	{5,12,13,17,18,19,20,24,25}
7	{ }
8	{4,11,12,22 }
9	{22}
10	{2,4,17,18 }
11	{7,18,19,22 }
12	{2,4,5,6,18 }
13	{1,2,6,8 }
14	{12,23 }
15	{1,7}
16	{7,22,24,25}
17	{ }
18	{1,11,12,13,19,24 }
19	{4,6,7,9,11,14,18}
20	{1,6,11,12,24 }
21	{1,2,5,6,7,22}
22	{3,6,8,11,21}
23	{ }
24	{1,10,13,16,18 }
25	{8,14,16,19 }

Chapter 5

CONCLUSION

In the presented work, we have shown that how existing patterns of friends & foes of an individual can be utilized to predict unknown links through inductive learning. Further, it is demonstrated that possible imbalance in the extended friends & foes network (FFN) of an individual can be avoided through social balance theory by appropriately choosing links from the set of all possible new links generated through the discovered rule-based classifier. The effectiveness of the proposed scheme is illustrated through experimental results. In this work, experiments are performed on a simulated dataset and we plan to conduct experiments on larger datasets obtained through survey/SNSs.

Future Work

We see many possibilities for future work based on the ideas presented in this paper. First of all, it would be interesting to explore different ways to incorporate trust-reputation [Bharadwaj and Al-Shamri, 2009] and trust-distrust [Anand and Bharadwaj, 2012] mechanism into our proposed scheme to further enhance its effectiveness. Secondly, exploitation of temporal [Kashoob and Caverlee, 2012] and spatial features [Guy et al., 2010] to make our system more accurate and efficient would also be an interesting future research. While looking towards another most promising research direction, group evolution discovery (GED) approach [Brodka et al., 2012] together with social-psychological balance theory may also have some important role in identifying and analyzing socially constructed groups within any structure of positive and negative arcs.

REFERENCES

Adafre, SF. and Rijke, MD. (2005), “Discovering missing links in wikipedia. In proceedings of *LinkKDD Workshop on Link Analysis and Link Detection*, Chicago, Illinois, USA, ACM.

Adamic, LA. and Adar, E. (2003), “Friends and neighbors on the Web”, *Social Networks 25(3),* pp.211–230, 2003

Adomavicius, G. and Tuzhilin, A. (2005), “Personalization Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions”, *IEEE Transaction on Knowledge and Data Engineering 17(6),* pp. 734–749.

Agarwal, V. and Bharadwaj, KK. (2011), “Trust-enhanced Recommendation of Friends in Web Based Social Networks Using Genetic Algorithms to Learn User Preferences”, *CCSEIT‘11 In: Proceedings of the First International Conference on Computer Science, Engineering and Information Technology, Springer-Verlag Berlin Heidelber*g, vol.204, pp. 476-490.

Ahmad, A., Borbora, Z. and Shrivastva, J. (2010), “Link Prediction Across Multiple Networks”, IEEE International Conference on Data Mining Workshops, pp. 911-918.

Anand D, Bharadwaj KK., “Pruning Trust-Distrust Network via Reliability and Risk Estimates for Quality Recommendations”, *Social Network Analysis and Mining,* Springer (In press).

Antal T., Krapivsky P. and Redner S. (2006), “Social balance on Networks: The Dynamics of Friendship and Enmity”, *Physica D 224(1),* pp. 130-136.

Bharadwa KK., Al-Shamri MYH. (2009), “Fuzzy Computational Models for Trust and Reputation Systems”, *Electronic Commerce Research and Applications, Elsevier.vol.8,* pp. 37-47.

Bian, L. and Holtzman, H. (2011), “Online Friend Recommendation through Personality Matching and Collaborative Filtering”, *UBICOMM 2011: The Fifth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*.

Bilenko, M., Mooney, RJ., Cohen, W., Ravikumar, P. and Fienberg, SE. (2003), “Adaptive name matching in information integration” *IEEE Intelligent Systems 18(5),* pp.16–23.

Boyd, DM. (2004), “Friendster and publicly articulated social networks”, In Proceedings of the *Conference on Human Factors and Computing Systems (CHI’04), Vienna, Austria*.

Boyd, DM. (2006), “Friends, Friendsters and Top 8: Writing community into being on social network sites”, *First Monday 11(12).*

Boyd, DM. and Ellison, NB. (2007), “Social Network Sites: Definition, History and Scholarship”, *Journal of Computer-Mediated Communication 13(1),* article 11.

Brodka, P., Saganowski, S. and Kazienko, P. (2012), “GED: the method for group evolution discovery in social networks”, *Social Network Analysis and Mining. Springer.*

Cartwright, D. and Harrary, F. (1956), “A generalization of Heider‘s Theory”, *Psychological Review* 63, pp. 277-292.

Chen, J., Dugan, C., Muller, M. and Guy, I. (2009), “Make New Friends, but Keep Old- Recommending People on Social Networking Sites”, *CHI‘09, In: Proceedings of the 27th International Conference on Human Factors in Computing Systems*.

Chiang KY., Natarajan N., Tewari A. and Dhillon IS. (2011), “Exploiting longer cycles for link prediction in signed networks”, *CIKM‘11: In Proceedings of the 20th ACM Conference on Information and Knowledge Management,* pp. 1157-1162.

Domingos, P. and Richardson, M. (2001), “Mining the network value of customers” In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining,* San Francisco, California, USA

Friedman, N., Getoor, L., Koller, D. and Pfeffer, A. (1999), “Learning probabilistic relational models”, In *IJCAI,* pp. 1300–1309.
Garcia, R. and Amatriain, X. (2010), “Weighted Content Based Methods for Recommending Connections in Online Social Networks”, *RECSYS’10 Barcelona, Spain, ACM.*

Getoor, L. (2003), “Link Mining: A New Data Mining Challenge”, *SIGKDD Explorations 5(1),* pp. 84-89.

Getoor, L. and Diehl, CP. (2005), “Link Mining: A Survey”, *SIGKDD Explorations 7(2),* pp. 3-12.

Golbeck, J. (2008), “The Dynamics of Web-based Social Networks: Membership, Relationships and Change”, In the Proceedings of the *International Sunbelt Social Network Conference (Sunbelt XXVIII),* st. Pete Beach, Florida.

Guha R., Kumar R., Raghavan P. and Tomkins A. (2004), “Propagation of Trust and Distrust”, *WWW’04, Proceedings of the 13th International Conference on World Wide Web, ACM,* New York, USA.

Guy, I., Jacovi, M., Perer, A., Ronen, I. and Uziel E. (2010), “Same Place, Same Things, Same People? Mining User Similarity on Social Media”, *CSCW‘10, In: Proceedings of the ACM Conference on Computer Supported Cooperative Work.*ACM New York, NY.

Hasan, M., Chaoji, V., Salem, S. and Zaki, M. (2006), “Link prediction using supervised learning”, *Workshop on Link Analysis, Counterterrorism and Security, SIAM Data Mining Conference,* Bethesda, Maryland.

Heckerman, D., Meek, C. and Koller, D. (2004), “Probabilistic models for relational data”, Technical report, Microsoft.

Heider, F. (1946), “Attitudes and Cognitive Organization”, *Journal of Psychology* 21, pp. 107-112.

Hsu, WH., King, AL., Paradesi, MSR., Pydimarri, T. and Weninger,T. (2006), "Collaborative and Structural Recommendation of Friends using Weblog-Based Social Network Analysis", *In AAAI Spring Symposia 2006 on Computational Approaches to Analysing Weblogs,* Stanford University, California, pp. 55-60.

Kashoob, S. and Caverlee, J. (2012), "Temporal Dynamics of Communities in Social Bookmarking Systems", *Social Network Analysis and Mining. Springer*.

Karkada, UH. (2009), "Friend Recommender System for Social Networks", *SI583 Term Paper, School of Information, University of Michigan.*

Kautz, H., Selman, B. and Shah, M. (1997), "Referral Web: Combining social networks and collaborative filtering" *Communications of the ACM 40(3),* pp. 63–65.

Khanafiah, D. and Situngkir, H. (2004), "Social Balance Theory", Technical Report WPN2004BFI, Dept. Computational Sociology, Bandung Fe Institute, Indonesia.

Kunegis, J., Lommatzsch, A. and Bauckhage, C. (2009), "The Slashdot Zoo: Mining a social network with negative edges", In *18th WWW'09*: Proceedings of the *18th international conference on World Wide Web,* pages 741–750.

Leskovec, J., Huttenlocher, D. and Kleinberg, J. (2010), " Signed networks in social media". In *CHI'10: Proceedings of the 28th international conference on Human factors in computing systems,* pp. 1361–1370.

Leskovec, J., Huttenlocher, D., and Kleinberg, J. (2010), "Predicting positive and negative links in online social networks." *In WWW'10: Proceedings of the 19th international conference on World Wide Web,* pages 641–650.

Liben-Nowell, D. and Kleinberg, J. (2007), "The Link Prediction Problem for Social Networks", *Journal of the American Society for Information Science and Technology 58(7),* pp. 1019-1031.

Mazid, MM., Ali SM, and Tickle, KS. (2010), "Improved C4.5 Algorithm for Rule Based Classification"**,** *AIKED'10: Proceedings of the 9th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases*.

McPherson, M., Smith-Lovin, L. and Cook, JM. (2001), "Birds of a feather: Homophily in social networks",Annual Review of Sociology 27, pp. 415–444.

Mislove, AE. (2009), "Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems", PhD thesis, Rice University, Houston, USA.

Monge, P. and Contractor, N. (2003), "Theories of Communication Networks", *Cambridge: Oxford University Press*

Quinlan, JR. (1986), "Induction of decision trees", *Machine Learning 1(1),* pp. 81–106.

Terveen, L. and Hill, W. (2001), "Beyond recommender systems: Helping people help each other*", In HCI in the New Millennium, J. Carroll, Ed. Addison-Wesley.*

Traag VA. and Bruggeman J. (2009), "Community detection in networks with positive and negative links", *PHYSICAL REVIEW E 80(3)*, 036115.

Wasserman, S. and Faust, K. (1994), "Social Network Analysis: Methods and Applications", Cambridge University Press, Cambridge.

Winkler, WE. (1994), "Advanced methods for record linkage", Technical report, Statistical Research Division, U.S. Census Bureau, Washington, DC.