STUDY ON THE PERFORMANCE OF ONTOLOGY BASED APPROACHES TO LINK PREDICTION IN SOCIAL NETWORKS AS THE NUMBER OF USERS INCREASES

by

SHRUTI PHANSE

B.E., Rajiv Gandhi Proudyogiki Vishwavidyalaya, India, 2008

A THESIS

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Computing and Information Sciences
College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas
2010

Approved by:

Major Professor
Doina Caragea

Copyright

Abstract

Recent advances in social network applications have resulted in millions of users joining such networks in the last few years. User data collected from social networks can be used for various data mining problems such as interest recommendations, friendship recommendations and many more. Social networks, in general, can be seen as a huge directed network graph representing users of the network (together with their information, e.g., user interests) and their interactions (also known as friendship links). Previous work [Hsu et al., 2007] on friendship link prediction has shown that graph features contain important predictive information. Furthermore, it has been shown that user interests can be used to improve link predictions, if they are organized into an explicitly or implicitly ontology [Haridas, 2009; Parimi, 2010]. However, the above mentioned previous studies have been performed using a small set of users in the social network *LiveJournal*. The goal of this work is to study the performance of the ontology based approach proposed in [Haridas, 2009], when number of users in the dataset is increased. More precisely, we study the performance of the approach in terms of performance for data sets consisting of 1000, 2000, 3000 and 4000 users. Our results show that the performance generally increases with the number of users. However, the problem becomes quickly intractable from a computation time point of view. As a part of our study, we also compare our results obtained using the ontology-based approach [Haridas, 2009] with results obtained with the LDA based approach in [Parimi, 2010], when such results are available.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to take the opportunity to thank all those people who in one way or another contributed and extended their valuable assistance in the preparation and completion of this thesis.

I owe my deepest gratitude to my major advisor, Dr. Doina Caragea, Assistant Professor in Computing and Information Sciences at KSU, for her support and encouragement throughout the completion of this project. Her valuable comments and suggestions at every step has inspired and helped me in overcoming obstacles faced during accomplishing my goals. I have learnt many things under her guidance in last two years and it gives me immense pleasure and honor to have had a major advisor like her.

It is a pleasure to thank my committee members, Dr. Gurdip Singh, Professor, Head of Computing and Information Sciences department at KSU and Dr. Torben Amtoft, Associate Professor in Computing and Information Sciences at KSU, for their valuable time and expertise comments, which have helped me to improve the quality of my work.

I would also like to thank the staff in Computing and Information Science department for their support and for providing me with the necessary resources needed to accomplish this project.

I wish to thank my friends for their valuable comments, suggestions and support at all the times.

I would like to acknowledge and thank Dr. Hsu and KDD group for sharing their *LiveJournal* social network data and findings from previous work.

Lastly, I would like to acknowledge Dr. Andresen for making available to us computational resources (specifically, the Beocat cluster) without which this research would not have been possible, and also for providing input on how to run scripts on Beocat, so that resources are optimally used.

This research was partially funded by a grant from National Science Foundation 0711396.

Dedication

I dedicate this thesis to my father, Dr. Deepak Phanse, and my mother, Dr. Neelima Phanse, who have always taught me that hard work and dedication helps us to achieve anything we aim for in life. Without their support and encouragement this would not have been possible.

Chapter 1

Introduction

1.1 Motivation & Problem Statement

Social network sites have been around since the mid-90's, but in recent years, these sites have exploded across the web. As a consequence the number of active users associated with the top ten social networking sites like *Facebook, Twitter, Linkedln* and *Bebo* have increased rapidly as reported by several online sources [1]. These online social networks give researchers unprecedented opportunities for data mining problems (like link prediction, interest prediction, tag recommendation, community formation, etc.).

A number of social network services like *LiveJournal, Orkut and Facebook* focus on user interactions [Fitzpatrick, 1999]. These services allow its users to list their interests and links to friends. In principle, *LiveJournal* social network can be represented as a directed graph structure, where nodes correspond to users of *LiveJournal* network and edges correspond to friendship links between these users. By directed we mean that, if a user A tags user B as a friend then it is not necessary that user B is also a friend of user A.

One of the basic computational problems associated with social networks is the *link prediction problem* [Taskar et al., 2003]. Link prediction can be described as a task to predict the existence of a friendship link from user A to user B in a social network.

Among others, two types of data are useful in predicting links in a social network: list of

[1] As of 2010, http://www.web-strategist.com/blog/2010/01/19/a-collection-of-social-network-stats-for-2010/

friends and list of interests of a user. More precisely, existence of a friendship link between two users depends on the knowledge of common interests and common friends they share. *Graph features* that can capture user friends, e.g. in-degree, out-degree, mutual friendship of two users, etc. have been used effectively at the task of link prediction [Hsu et al., 2007]. *Interest features* generated by making use of the information gained from the shared interests of two users have shown promising results at the task of link prediction problem. However, the large number of interests makes it hard to use this information effectively. One way to address this problem and thus reduce the dimensionality of interest feature vectors is to group interests into an ontology. Previous work [Bahirwani, 2008; Haridas, 2009] on a small *LiveJournal* dataset consisting of 1000 users have suggested that combining graph based features with interest based features (generated in the presence of an interest ontology) results in better performance at the task of link prediction in *LiveJournal* social network as compared with using interest features themselves [Hsu et al., 2007]. However, it is not clear if ontologies can be effectively constructed from larger data sets or if the performance of the resulting classifiers improves with the size of the data sets. Therefore, the main objective of this work is to study the variation in the performance of the classifiers at the task of link prediction when presented with graph-based features alone, interest-based features alone and graph plus interest based features generated using increasingly large data sets. Specifically, in this thesis, we have considered four subsets of the *LiveJournal* data consisting of 1000, 2000, 3000 and 4000 users, respectively. Furthermore, a recent work presented by Parimi [2010] has proposed an alternative, Latent Dirichlet Allocation (LDA), based approach to the interest dimensionality reduction problem. Thus, another objective of this work is to compare the ontology based/proposed approach used in this work with the LDA based approach.

1.2 Overview of the Proposed Approach

In this thesis we use data extracted from *LiveJournal* social network, which contains detailed information about the users of this network. The *LiveJournal* data consists of approximately 37,989 users, 371,133 declared friendship links and 2,151,873 declared interests. As mentioned above our objective is to study the performance of the ontology approach using varying size subsets of the *LiveJournal* data. More precisely, we study the performance of various classifiers at the task of predicting friendship links between users of *LiveJournal* social network subsets of size 1000, 2000, 3000 and 4000 users. Furthermore, we compare the performance of the classifiers when presented with features constructed using the ontology approach and when presented with features constructed using the LDA topic modeling approach, when the corresponding LDA results are available.

We use the approach proposed in [Haridas, 2009] to construct an interest ontology based on the complete category links for user interests fetched from *Google Directory* (we have used *Google Directory* instead of *Directory Mozilla/Open Directory Project*). *Google Directory* integrates Open Directory Project (ODP) pages and Google's search technology to find information on the Web more efficiently. We use Google's AJAX search API in order to get definitions of user interests for the construction of the interest ontology.

We analyze the performance of friendship link prediction problem in the *LiveJournal* social network, when graph based features are used alone; interest based features are used alone; and when graph based features are used in combination with interest-based features. *Graph based features* are represented by *in-degree, out-degree, mutual friends* and *backward distance*. These features are calculated by considering the graph structure of *LiveJournal* social network. The *interest based features* (represented as eight numeric features suggested by Aljandal et al. [2008]) are computed by considering the interests at different levels of abstraction in the interest ontology. We calculate *graph based features* and *interest based features* for all four datasets considered for experimentation. Performance is evaluated on

the basis of how well a machine learning classifier performs at the task of link prediction.

Hsu et al. [2007], Bahirwani [2008] and Haridas [2009] have suggested that *graph based features* in combination with *interest based features* give better results than *graph based features* or *interest based features* used alone. On the same note, we want to see if this hypothesis holds for the larger datasets considered in this thesis. Furthermore, we want to see if the features constructed in the presence of the interest ontology give better performance as compared to the features constructed when LDA approach was used as suggested by Parimi [2010].

The rest of the thesis is organized as follows: Chapter 2 describes other work related to the problem addressed. In Chapter 3 describes the ontology engineering approach considered in this thesis and method used for feature construction. Chapter 4 and Chapter 5 discuss the experiments performed and the results obtained when predicting friend relationships between the users, respectively. We conclude and present ideas for future work in Chapter 6.

Chapter 2

Related Work

Social Networks have become popular in past few years, as a result Social Networks has attracted a number of researchers for various data mining opportunities. This chapter reviews a number of works done in past, which are related to the objectives of this thesis.

2.1 Link Prediction in Social Networks

Getoor [2003] have addressed several link prediction problems such as link-based classification, link-based cluster analysis, identifying link type, predicting link strength and cardinality. Taskar et al. [2003] have focused on predicting the existence of links and type of links between entities of relational dataset. They used a *Relational Markov Network* (RMN) framework in order to define the joint probabilistic model over the entire link graph, which exploits the attributes and links of the entities in the network. The link graph reflects interactions between the entities in the domain, interactions that help in the link prediction task. For this task the RMN algorithm used the probabilistic patterns over subgraph structures. Their work suggested a significant improvement in the classification task by making use of RMNs classification approach and introduction of subgraph patterns over link labels.

Social networks are highly dynamic in nature, they grow and change quickly over time. Work presented in [Nowell and Kleinbergz, 2004] addresses link prediction problem keeping in mind the dynamic nature of social networks: Given a snapshot of a social network at time t, they try to predict the edges that will be added to the network during the interval

from time t to a given future time t‘. Experimental results suggested that information about future interactions could be extracted from network topology alone. Recent studies presented in [Song et al., 2009] exploited proximity measures (which defines the closeness or similarity between nodes in a social network) in order to predict links, which can be added to the current snapshot of a social network.

Work presented in [Patil, 2009] addresses the problem of predicting friendships between actors in a social network by making use of user's socio-demographic attributes. Patil [2009] makes use of principle of homophily in order to exploit social relationships among similar individuals. Similar individuals may share same culture, language, hobby, social institution like workplace, school etc. Patil [2009] approach evaluates most important attributes that are associated with friendship formation by making use of chi-square approach. Conclusions from [Patil, 2009] suggested that social links between two individuals can be predicted by exploiting the user's demographic information.

Some of the findings presented in [Tang and Liu, 2010], explored new challenges for more effective graph mining techniques. They suggested that, large-scale networks share some common patterns that are not noticeable in small-scale networks. In particular, they discussed graph mining applications to community detection in order to demonstrate strong community structures in large-scale networks. Community effect can be seen as, a group of people who tend to interact with each other more often than those outside the group. The work presented in [Tang and Liu, 2010] categorized a number of representative graph mining approaches and evaluation strategies for community detection.

Other than social networks, link prediction problem can be associated with various real networks. Work presented in [Lu and Zhou, 2010] estimated the likelihood of existence of links in weighted networks. Conclusions from this work suggested that links with small weights played an important role in link prediction.

2.2 Ontology Construction & Usage

Some approaches to constructing ontologies relies on Web directories, which offer a sophisticated way of browsing through Web. Web directory maintain the Web organized in subject hierarchies. Some of the popular Web Directories are Google Directory [1], Yahoo! Directory [2] and Open Directory Project [3]. With the growth of available data on the Web in the form of information from social networking sites, Web documents and articles have attracted researchers towards innovative data mining techniques.

Grobelnik and Mladenič [2005] presented an approach to classify web documents into an existing topic ontology. They have presented an approach for constructing an ontology from a stream of documents. In order to classify a document they have used document content as well as the information on the Web page context obtained from the link structure of the Web. Their work showed the usefulness of *Directory Mozilla* (DMoz), i.e. a large topic ontology in classifying web documents.

Ceci and Malerba [2007] have explored a way to improve the efficiency and efficacy of text classification by making use of hierarchical structures, such as Yahoo! and DMoz. Text categorization mainly focuses on classifying text documents into a set of categories without considering structural relationships among them. Ceci and Malerba [2007] used existing hierarchical structure employed by many Internet directories and involved the hierarchy of categories in all phases of text categorization. Conclusions from this work suggest that large collections of documents can be organized in categories. Furthermore, the hierarchical approach had two advantages: it showed gain in the efficiency and reduction in classification errors.

Grobelnik et al. [2006] have constructed an ontology over a stream of documents. The approach in this work showed that by using DMoz as an existing topic hierarchy, concepts and relations could be formed into an ontological structure. These ontologies help in the

[1] http://dir.google.com
[2] http://yahoo.com
[3] http://dmoz.org

process of data understanding and analysis. This approach was efficient, scalable, and was able to process large quantities of data including thousands of documents. Work presented in [Burger, 2010] developed a classification-method, called Virtus, which has the ability to classify documents (like text files) into an ontology by making use of their metadata. Thus, suggesting usefulness of ontologies for classifying documents efficiently.

2.3 Link Prediction using Ontologies

Researchers have explored in the past, the usefulness of constructing ontologies over data and have shown that such ontologies can provide a better understanding of the data. In social networks, ontologies can be used to provide a semantic organization of the knowledge/data available in such networks. Some of the major applications of ontologies can be seen in problems such as classifying user interests in social networks, text classification, document classification, etc. A number of approaches make use of clustering algorithms over collection of terms or documents in order to engineer hierarchical ontology. Bahirwani [2008], Haridas [2009] and Parimi [2010] and Hsu et al. [2007] have reported their findings related to the task of predicting friendship links in *LiveJournal* social network.

Hsu et al. [2007] have investigated the problem of link recommendation over a small subset of *LiveJournal* social network. Their approach uses the network structure and user profile data to recommend links. More specifically, they have used a hybrid system that combines analysis of link structure with analysis of content, such as shared interests. Their results suggest that graph and annotated features combined resulted in better recommendations than interest-based or simple graph-based recommendations.

The above approach was extended by Bahirwani [2008] to make better use of user interests. He proposed to organize user interests into an ontology. Bahirwani [2008] constructed an ontology by fetching definitions of user interests from three different online sources namely WordNet-Online, Internet Movie Database (IMDB) and Amazon Associates Web Services (AWS). Interests can have more than one definition and these definitions are seen as in-

stances. Similarity between two instances is defined as the number of common terms describing the instances. This similarity was evaluated as the dot product between vectors representing these instances. These instances were grouped together into a hierarchical ontology using a hierarchical agglomerative algorithm. The approach described by Bahirwani [2008] suffered from some limitations, but it showed improvement over the Hsu et al. [2007] approach at the task of predicting friendship links.

Haridas [2009] used more comprehensive knowledge bases as compared to Bahirwani [2008] (in particular, *Wikipedia* (Wiki) and *Directory Mozilla* (DMoz)) in order to obtain definitions for interests belonging to a wider variety of domains. He proposed three different approaches to built hierarchical ontologies over user interests. In their first approach, he obtained definitions of user interests from *Wikipedia*. Furthermore, he used *Latent Semantic Analysis* (LSA) to calculate the similarity between interest documents. In the second and third approaches he exploited the knowledge from existing hierarchies like Wikipedia Category Graph (WCG) and Directory Mozilla (DMoz), respectively. As reported by Haridas [2009], the third approach proved effective for building the ontology over user interests and was further used to construct interest-based features.

A new dimension to the task of link prediction was added by Parimi [2010]. He used topic modeling techniques (specifically, Latent Dirichlet Allocation (LDA)) on interests specified by the users of *LiveJournal* social network to address the link prediction problem. LDA is a machine learning technique used to uncover latent structure in text in the social network problem (text here corresponds to interests of users of *LiveJournal* social network). Each user in the dataset was seen as a document and the content of a document corresponded to his/her interests. Thus, each document was treated as a mixture of topics and each topic in turn was treated as a mixture of words. By using LDA to group interests, Parimi [2010] does not construct an ontology explicitly, but implicitly simulates an ontology by varying the number of latent topics to be identified. Conclusions from this work suggested an improvement in the performance of link prediction problem with increase in the number

of users considered in the *LiveJournal* social network dataset.

2.4 Work in this Thesis

Our work builds up on and extends the work presented by Bahirwani [2008] and Haridas [2009] along several directions. (1) Haridas [2009]'s work used DMoz to built the ontology over user interests of *LiveJournal* social network, while in this thesis we have used *Google Directory* to extract interest definitions. As mentioned before *Google Directory* is a better and enhanced way to navigate through DMoz. (2) Hsu et al. [2007], Haridas [2009] and Bahirwani [2008] have used a small subset consisting of 1000 users of *LiveJournal* social network and suggested that combination of graph-based features and interest-based features (in the presence of the interest ontology) give better performance at the link prediction task. We study the variation in performance of various classifiers over larger subset of data, consisting of 2000, 3000 and 4000 users. Better performance of the classifiers is expected, because the graph becomes more connected as we scale up the number of users. (3) Recent work reported by Parimi [2010] suggested an innovative way of simulating an implicit ontology using LDA topic modeling technique and showed an improvement in link prediction problem in social networks. We compare our proposed methodology with the approach presented by Parimi [2010].

Although, the work presented in this thesis explores and extends previous work [Hsu et al., 2007], [Bahirwani, 2008], [Haridas, 2009] and [Parimi, 2010] on the link prediction problem in social networks, to the best of our knowledge contributions of this thesis have not been explored before.

Chapter 3

Ontology Engineering Approach & Feature Construction

In this Chapter we describe the ontology engineering approach considered in this thesis. In order to construct the interest ontology using the user interests from the datasets considered in this work, we use the methodology introduced in [Haridas, 2009]. To evaluate the performance of the machine learning classifiers using the four datasets at the task of link prediction in *LiveJournal* social network we generate graph-based features, interest-based features using the interest ontology, and interest plus graph based features. Graph-based features generated in this work are constructed using the approach introduced in [Bahirwani, 2008]. Similarly, we generate interest-based features using the interest ontology following the approach mentioned in [Haridas, 2009].

This Chapter is organized as follows: in Section 3.1 we discuss the ontology engineering approach; in Section 3.2 we describe approach followed to generate interest-based features; in Section 3.3 we describe the approach followed to generate graph-based features and in Section 3.4 we show how we combine graph and interest based features.

3.1 Ontology Engineering Approach

For ontology construction, we follow the third approach presented by Haridas [2009]. Details regarding the construction of ontologies using this third approach are as follows: In the first

Table 3.1: *Details of the ontologies generated using 4 subsets of the LiveJournal social network.*

Datasets	Nr of Interests	Nr of Levels	Nr of Leaf Nodes	Nr of Internal Nodes
1,000	14,430	12	20,147	17,695
2,000	21,800	12	41,053	25,831
3,000	26,200	12	72,457	28,641
4,000	29,900	12	108,849	30,724

step, we fetch interests definitions or category links from *Google Directory*. For this purpose we have used Google's search AJAX API. An interest can have multiple definitions. The category links extracted help in ranking the categories in the hierarchical ontology. We have generated four such ontologies based on user interests from four datasets considered in this thesis. The ontologies constructed above are hierarchial in structure and are stored in a MySQL database. Details regarding these features for the ontologies generated can be seen in Table 3.1.

3.2 Computing Interest Based Features

The Link prediction problem in a social network can be defined as the problem of predicting the existence of a friendship link between a given pair of users $\langle A, B \rangle$. Given a pair of users, user A and user B, the interests of the two users and the common interests between them, we can address the above problem by computing interest based numerical features. The intuition is, if two users share many interests, then it is likely that they are friends regardless of what these interests are. Similarly, if two users share a very rare interest, then they might be friends too. Several interest based features can be derived, which capture this intuition and help us to address the link prediction problem.

We derive interest-based features using the methodology described in [Aljandal et al., 2008], where the author derived eight such numerical features by measuring the interestingness between the two users having some common interests. Each interest in *LiveJournal* data has an item set, elements of this item set are the users having this interest. Association

rules of the form $A \longrightarrow B$ are formed from these item sets (where A and B are users of this data). These association rules help in constructing eight objective measures of rule interestingness, which further help in deriving interest-based numerical features. We derive interest features for the four datasets considered in this work. These interest-based features are derived by making use of the ontologies derived over the four datasets used in this work. Interests of users of *LiveJournal* social network are viewed at different levels of abstraction and experiments conducted over these features help in revealing the levels in the ontology which give the best performance of the classifiers.

3.3 Computing Graph Based Features

Social networks like *LiveJournal* can be represented as a graph with nodes corresponding to users and edges corresponding to the friendship links between the users. Such features include in-degree, out-degree, mutual friendship and backward deleted distance. A total of nine graph-based features are derived [Hsu et al., 2007] over the four data sets of *LiveJournal* data considered in this thesis.

3.4 Computing Interest and Graph Based Features

In order to construct interest plus graph based features we combine the interest features generated at a particular level of the interest ontology with the graph based features. The methods described in Section 3.2 and 3.3 are used to generate the interest based features and graph based features, respectively.

Chapter 4

Experimental Setup

The experiments designed and performed in this thesis help us to address a number of questions raised. These questions are: Does the combination of *Graph-based Features* and *Interest-based Features* give better results than when *Graph-based Features* or *Interest-based Features* are used alone, when working with data sets containing more than 1000 users? How do the classifiers perform over the data sets used in this work? Out of the four classifiers, which one performs the best? How does the ontology approach considered in this thesis perform when the number of users is increased? Does the LDA based approach mentioned in [Parimi, 2010] perform better than the ontology approach considered in this thesis at the task of predicting friendship links in *LiveJournal* social network?

To address the questions above, we use four classifiers. Performance of these classifiers was measured over the four data sets considered in this work, when used with interest based features constructed using interest ontology, alone, or in combination with graph based features, generated from the corresponding data sets network graph structure. The performance of these learning algorithms can be measured through some performance measures like *Recall, Precision, F-Measure, ROC Area, Accuracy* etc [Mitchell, 1997]. In this work, we use the area under the *Receiver Operating Characteristic* (ROC) curve. We hypothesize that the performance of the classifiers improves as we increase the graph size, i.e. as we scale up the number of users. Experiments described in this Chapter are meant to help us test this hypothesis.

The Chapter is organized as follows: in Section 4.1, we explain the details of the data sets considered for the experiments conducted. In Section 4.2 we describe our experimental design and Section 4.3 presents the Machine Learning Algorithms used and the need for using filters for balancing the training data sets.

4.1 Dataset

In this thesis, all experiments are performed over *LiveJournal* social network data. Usually, social networks like *LiveJournal* consists of thousands of users. The data that was available to us consists of approximately 38,000 users. Details of this data are provided in Table 4.1.

Interests expressed by users of *LiveJournal* social network fall under a wide variety of domains, to name a few these domains includes *Movies, Books, Current Issues*, *Sports* etc. For experimentation purpose we generate four subsets of the data. These subsets consists of 1000, 2000, 3000 and 4000 users and their corresponding declared interests and friendship links, respectively. It is important to pre-process the data before using it for experiments as it contains symbols and foreign language characters. Since, one of the goals of our thesis is to compare our results with those presented by Parimi [2010], we follow the exact procedure to clean the interests as mentioned by Parimi [2010]. Thus, during pre-processing step we remove symbols, foreign language characters and all those interests which have frequency less than 5. We also clean and remove all *bad users* from our subsets. These users are the ones who do not have any declared interests and who have zero in-degree and out-degree. Users left after pre-processing are addressed as *good users*. This procedure is repeated for all four subsets of data considered.

Ontologies are constructed by using these four pre-processed subsets of data. Interest

Table 4.1: *Details of the complete LiveJournal data set available*

Total number of Users	Total number of Interests	Total number of Friendship Links
37,989	2,151,873	371,133

Table 4.2: *Details of the subsets of LiveJournal Data used for Experimentation*

Total number of Users	Good Users	Interests	Interest-Instances	Friendship Links
1,000	801	14,430	37,800	4,452
2,000	1,602	21,800	66,800	9,020
3,000	2,440	26,200	101,000	14,780
4,000	3,240	29,900	139,000	22,000

definitions are fetched from *Google Directory* for a particular data set and organized into corresponding interest ontologies. Complete details of these four pre-processed data subsets with the number of good users, number of interest-instances, number of friendship links and number of interests are provided in Table 4.2. After pre-processing only 801, 1602, 2440 and 3240 good users are left in the four data sets respectively. The number of declared friendship links shown in the Table 4.2 is out of $n \times n$ possible links in an undirected graph with n nodes, i.e. *good users*. We assume that all the declared friendship links are positive instances and non-declared friendship links are negative instances. The ratio of positive instances to negative instances is 1:143 and is kept unchanged in the test data sets. This makes our data sets highly skewed (unbalanced) toward the negative class.

Each data set is further divided in two parts such that two-third of the good users come in training set and the remaining one-third go in the test set. We generate five such random splits for each subset of the data. Partition of users in two-third and one-third sets is done such that the two sets are independent. By independent we mean to say that, we do not consider friendship links that go across train and test set. *Graph-based features*, *Interest-based features* and combination of *Graph-based* and *Interest-based* features are generated over five random splits and over four subsets of the data.

4.2 Experiments

The experiments that are performed to evaluate the ontology based approach at the task of predicting friendship links using the interest ontology are listed below:

1. *Interest-based numeric features with the ontology:*

In the first experiment we construct the interest-based features for 1000 user data set. We generate these features at different levels of abstraction in the ontology as explained in Chapter 3 Section 3.2. Performance of all classifiers are learnt on the training set consisting of 1000 users, represented using the interest based features and evaluated on a separate test set. We refer to this experiment as Experiment 1.

2. *Graph-based features:*

   The second experiment addresses the problem of predicting friendships by exploiting the graph-based features described in Chapter 3 Section 3.3. In particular, graph-based features are generated when 10%, 25 % and 50% of links are known in the test set. Performance of all the classifiers is evaluated when classifiers are learned on the training set consisting of 1000 users, represented using the graph based features and evaluated on separate test set. We refer to this experiment as Experiment 2.

3. *Graph-based features and interest-based numerical features with the ontology:*

   In the third experiment we use the graph-based and interest-based numeric features to predict friendship links. These features are generated for 1000 user dataset in this experiment. Particulary, graph-based features generated using 10%, 25% and 50% links known are combined with interest-based features generated at different levels of abstraction in the ontology. With the use of these features we expect to see an improvement in the performance of classifiers as compared to results from Experiment 1 and Experiment 2. We refer this experiment as Experiment 3.

Since, the data sets used in this thesis are highly unbalanced we proposed to balance these data sets by using undersampling technique i.e. SpreadSubsample (mentioned in Section 4.3.1). In each experiment above we first balance the train data sets by using the SpreadSubsample filter with 1:1 spread and 2:1 spread between the negative and positive instances. Thus, we refer the Experiment 1 with 1:1 spread as Experiment 1(a) and with 2:1 spread as 1(b). Similarly, Experiment 2 and Experiment 3 with 1:1 spread are referred

as Experiment 2(a) and Experiment 3(a), respectively, and with 2:1 spread are referred as Experiment 2(b) and Experiment 3(b), respectively. These experiments are performed over four subsets of data described in Section 4.1. Experiments on 2000 users dataset are referred as Experiment 4, Experiment 5 and Experiment 6; experiments on 3000 users dataset are addresses as Experiment 7, Experiment 8 and Experiment 9; and experiments on 4000 users dataset are labled as Experiment 10, Experiment 11 and Experiment 12.

Furthermore, under Experiment 13 and Experiment 14 we compare the results obtain for interest-based features and interest+graph based features for 1000 users data set with the results obtained in the work presented by Parimi [2010]. These 14 experiments mentioned here will help us in answering the questions raised above. Results from the above experiments are reported in Chapter 5 and conclusions derived from them are discussed in Chapter 6.1.

4.3 Machine Learning Algorithms Used

The classifiers considered while performing the experiments are mentioned below. Implementations of these classifiers are provided by the WEKA data mining software [Witten et al., 1999].

- Random Forest (RF)
- J48 Decision Tree (J48)
- Logistic Regression (LR)
- Support Vector Machines (SVM) with *built logistic model* option enabled

The performance of each classifier is measured by the area under the *Receiver Operating Characteristic* (ROC) curve. The ROC depicts the tradeoff between *true positive rate* vs. *false positive rate*. The value of the area under the ROC (called AUC) ranges between 0 and 1. A higher AUC value of the ROC implies a better performance by the classifier at

the task of link prediction, while the opposite holds for lower values. Results are reported as average of the AUC values over five random splits of the data sets.

4.3.1 SpreadSubsample Filter for Undersampling

The performance of machine learning classifiers is highly dependent on the training set that it is supplied with, as the classifier learns from the training set. For our problem, the training set consists of instances belonging to a majority class and a minority class. The datasets used in experimentation are highly skewed towards the majority class. When a classifier learns from a training distribution which is biased strongly in the favor of the majority class it tends to predict the majority class in the test set and thus it performs poorly at classifying the minority class. Thus, the prediction capability of a classifier tends to be biased towards the majority class in the training set as studied by Weiss and Provost [2001]. In order to improve the performance of the classifier, the class distribution of the training set is modified. We make our data sets balanced by using an undersampling technique. For this purpose we use the SpreadSubsample filter provided by the WEKA data mining software [Witten et al., 1999]. It produces a random subsample of the input data set. SpreadSubsample gives an option to specify the maximum spread between the majority and minority class. For experimentation purpose we used 2:1 spread and 1:1 spread between the majority and minority classes.

Chapter 5

Results

In this chapter we discuss the results obtained by running the experiments mentioned in Chapter 4. To obtain results, classifiers are run over the interest based and graph based features generated for datasets containing 1000, 2000, 3000 and 4000 users of *LiveJournal* social network. Previous work have suggested that the performance of the classifiers improved when graph-based features are used in combination with interest-based features. Our aim is to study the performance of these classifiers when the number of users in the dataset increases. Results obtained from these classifier do indicate an improvement in the performance, when graph-based features are used in combination with interest-based features over the four datasets considered here. Performance of the classifiers also improves when the number of users in the dataset increases.

This chapter is organized in three sections: in Section 5.1 we report results obtained for all datasets considered in this work, in Section 5.2 we discuss the performance of the ontology approach when the number of users is increased in the datasets, and in Section 5.3 we compare the results obtained from 1000 user dataset with the results obtained in the work presented by Parimi [2010].

5.1 Predicting Friendship Links

In this section we investigate the results obtained from the classifiers at the task of predicting friendship links between the users of *LiveJournal* social network. Furthermore, we also

explore the usefulness of considering various levels in the interest ontology, while constructing interest based features. Performance of all the classifiers at the task of predicting friendship links are reported as AUC values, averaged over 5 different partitions of the user graph. This section is further divided into four sub-sections. Results from experiment 1, 2 and 3 are reported in sub-section 5.1.1; sub-section 5.1.2 describes results from experiment 4, 5 and 6; sub-section 5.1.3 shows results from experiment 7, 8 and 9; and sub-section 5.1.4 describes results from experiment 10, 11 and 12, respectively.

5.1.1 Results from 1000 user dataset

The AUC values reported by running experiments 1(a), 2(a) and 3(a) using Logistic Regression with 1:1 spread, Support Vector Machine with 1:1 spread, Random Forest with 1:1 spread and J48 Decision Trees with 1:1 spread classifiers are reported in Table 5.1. The AUC values reported by running experiments 1(b), 2(b) and 3(b) using Logistic Regression with 2:1 spread, Support Vector Machine with 2:1 spread, Random Forest with 2:1 spread and J48 Decision Trees with 2:1 spread classifiers are reported in Table 5.2. In above tables, we compare the AUC values obtained by using interest-based features, graph-based features and interest plus graph based features. Interest features are generated over different ontology levels, graph features are generated using a certain percentage of links known and interest plus graph features are generated over different ontology levels using a certain percentage of links known. The ontology level at which the highest AUC value is observed for interest and interest plus graph features is mentioned in the brackets i.e. '()'. We have highlighted (in **Bold-Faced**) the highest AUC values obtained for all four classifiers in Table 5.1 and 5.2.

Tables 5.1 and 5.2 report the results obtained from classifiers with 1:1 spread and 2:1 spread using features generated for 1000 user dataset, respectively. Graphs are generated by plotting AUC values obtained using graph-based features, interest-based features and interest+graph based features for 1,000 users of the *LiveJournal* social network against different levels in ontology (constructed over interests of 1,000 users). Figure 5.1 and 5.2

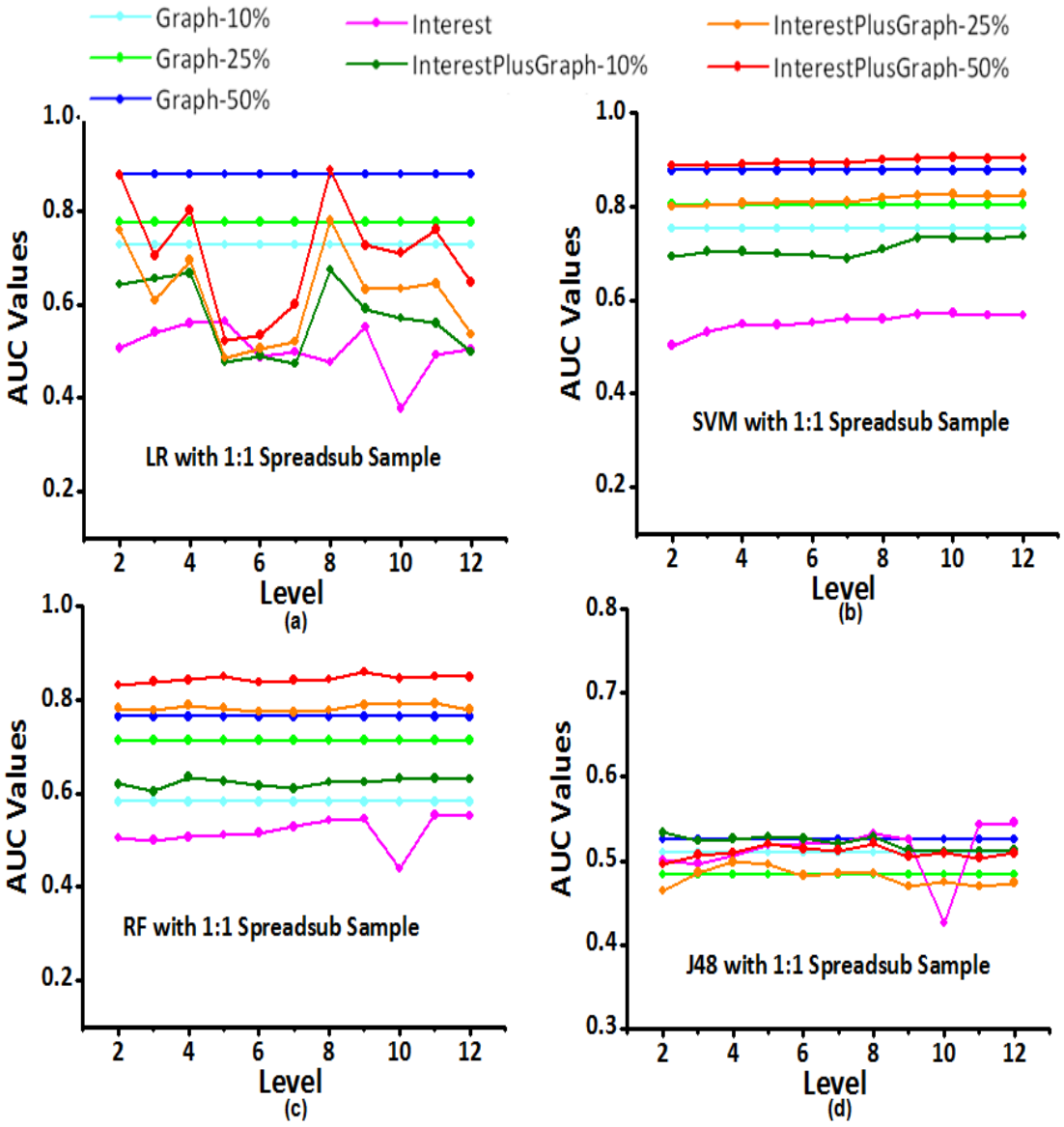

Figure 5.1: *Graphs obtained by plotting AUC values against the number of levels in the interest ontology, (a) Logistic Regression with 1:1 spread, (b) Support Vector Machine classifier with 1:1 spread, (c) Random Forest with 1:1 spread, and (d) J48 classifier with 1:1 spread for 1,000 users dataset.*

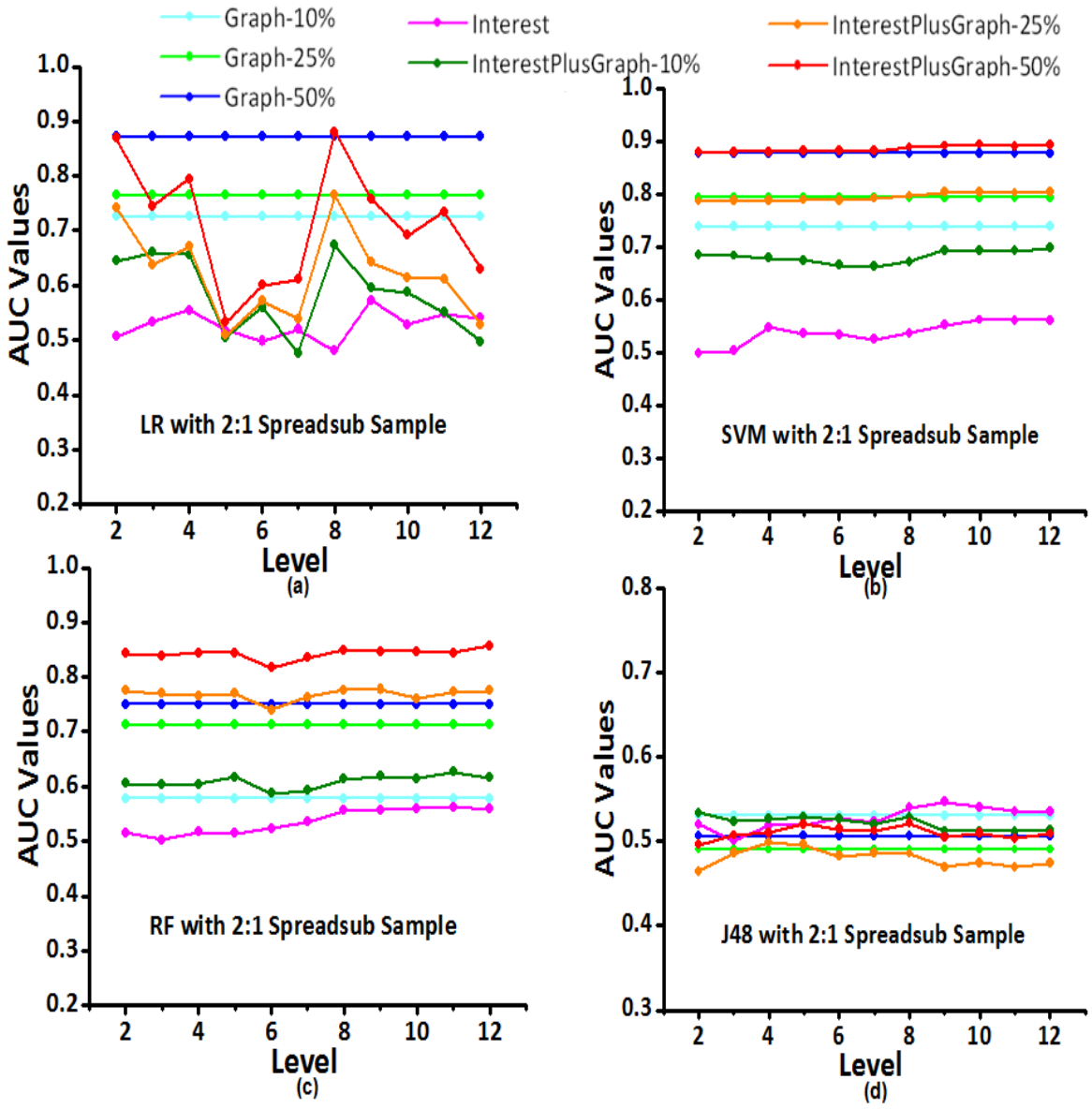

Figure 5.2: *Graphs obtained by plotting AUC values against the number of levels in the interest ontology, (a) Logistic Regression with 2:1 spread, (b) Support Vector Machine classifier with 2:1 spread, (c) Random Forest with 2:1 spread, and (d) J48 classifier with 2:1 spread for 1,000 users dataset.*

Table 5.1: *AUC values for Logistic Regression (LR), Random Forests (RF), Support Vector Machines (SVM) and J48 Decision Trees (J48) classifiers with interest, graph and interest+graph based features using 1:1 spread for the 1,000 user dataset. Here we have assumed that k% links are known in the test set, where k is 10, 25 and 50, respectively. K% known links are used to construct graph features and interest+graph features.*

Exp	Features	Logistic	SVM	RandomForest	J48
1(a)	Interest	0.564(4)	0.5718(11)	0.5536(11)	**0.5454(12)**
2(a)	Graph 10%	**0.7258**	**0.7418**	0.5698	0.5102
3(a)	Graph+Interest 10%	0.6672(4)	0.7376(12)	**0.6346(4)**	0.5288(8)
2(a)	Graph 25%	0.7624	0.7924	0.7036	0.4838
3(a)	Graph+Interest 25%	**0.78(8)**	**0.8258(12)**	**0.7926(11)**	0.4988(4)
2(a)	Graph 50%	0.8538	0.8624	0.7798	0.5254
3(a)	Graph+Interest 50%	**0.8878(8)**	**0.904(12)**	**0.8498(11)**	0.521(8)

Table 5.2: *AUC values for Logistic Regression (LR), Random Forests (RF), Support Vector Machines (SVM) and J48 Decision Trees (J48) classifiers with interest, graph and interest+graph based features using 2:1 spread for the 1,000 user dataset. Here we have assumed that k% links are known in the test set, where k is 10, 25 and 50, respectively. K% known links are used to construct graph features and interest+graph features.*

Exp	Features	Logistic	SVM	RandomForest	J48
1(b)	Interest	0.5728(9)	0.5626(10)	0.5628(11)	**0.5352(11)**
2(b)	Graph 10%	**0.74**	**0.7738**	0.578	0.5302
3(b)	Graph+Interest 10%	0.6728(8)	0.6982(12)	**0.6178(9)**	0.5338(2)
2(b)	Graph 25%	**0.7684**	**0.8104**	0.7106	0.4902
3(b)	Graph+Interest 25%	0.7654(8)	0.8052(12)	**0.7772(9)**	0.4988(4)
2(b)	Graph 50%	0.8526	0.8692	0.8008	0.5062
3(b)	Graph+Interest 50%	**0.8808(8)**	**0.8948(10)**	**0.8576(12)**	0.521(8)

plots these graphs for all four classifier with 1:1 and 2:1 spread, respectively. Studying the performance of Logistic Regression classifier (in Fig. 5.1 (a) and 5.2 (a)) we can observe that graph features alone outperforms interest features alone. Also, graph features with 10% links known are better than interest+graph features with 10% links known. However, highest AUC value is obtained for interest+graph features with 25% and 50% links know as opposed to graph features with 25% and 50% links known, but these values are not consistent across all levels of ontology. However, interest+graph features with 25% and 50% give better performance at all levels of ontology as compared to graph features with 25% and 50% links

known for Support Vector Machine classifier (can be seen in Fig. 5.1 (b) and 5.2 (b)), while, graph features with 10% known links outperforms interest+graph features with 10% known links across all levels of ontology for Support Vector Machine classifier. However, this is not the case with Random Forest. Interest+graph features with 10%, 25% and 50% known links give better performance across all levels of ontology as compared to graph features with 10%, 25% and 50% known links for Random Forest classifiers as seen in Figure 5.1 (c) and 5.2 (c). On the other hand, performance of J48 classifier is very inconsistent and also poor as compared to other classifiers. J48 classifier didn't prove to be a good classifier for 1000 user dataset, refer Fig. 5.1 (d) and 5.2 (d)).

5.1.2 Results from 2000 user dataset

Studying Table 5.3 and Table 5.4 it can be seen that in terms of AUC values graph features with 10% known links outperformed interest+graph features with 10% links known for Logistic Regression with spread 1:1 and 2:1. However, interest+graph features with 25% and 50% links known outperformed graph features with 25% and 50% links known for Logistic Regression classifier with 1:1 spread and 2:1 spread. Furthermore, graph feature with 10%, 25% and 50% links known outperformed interest+graph features with 10%, 25% and 50% links known for Support Vector Machine with spread 2:1 and the opposite for 1:1 spread. However, interest+graph features with 10%, 25% and 50% links known outperformed graph features with 10%, 25% and 50% links known for Random Forest and J48 classifier with 1:1 spread and 2:1 spread (Table 5.3 and 5.4).

Comparing the results obtained for 2,000 users dataset and 1,000 user dataset we see a lot of difference in the performance of the classifiers. There is a substantial improvement in the performance of J48 classifier with 1:1 and 2:1 spread when presented with interest features, graph features with 10%, 25%, 50% links known and interest+graph features with 10%, 25%, 50% links known (Table 5.1 and 5.2, 5.3 and 5.4). Improvement is also shown by Random Forest classifier with 1:1 spread and 2:1 spread for graph features with 50% links

Table 5.3: *AUC values for Logistic Regression (LR), Random Forests (RF), Support Vector Machines (SVM) and J48 Decision Trees (J48) classifiers with interest, graph and interest+graph based features using 1:1 spread for the 2,000 user dataset. Here we have assumed that k% links are known in the test set, where k is 10, 25 and 50, respectively. K% known links are used to construct graph features and interest+graph features.*

Exp	Features	Logistic	SVM	RandomForest	J48
4(a)	Interest	0.6042(10)	0.6136(9)	0.5848(10)	0.5862(11)
5(a)	Graph 10%	**0.6798**	0.5264	0.5786	0.5414
6(a)	Graph+Interest 10%	0.6722(9)	**0.6638(12)**	**0.6722(9)**	**0.5604(3)**
5(a)	Graph 25%	0.7508	0.622	0.681	0.6018
6(a)	Graph+Interest 25%	**0.7584(2)**	**0.757(12)**	**0.7584(2)**	**0.6374(3)**
5(a)	Graph 50%	0.855	0.7736	0.8012	0.7008
6(a)	Graph+Interest 50%	**0.8718(2)**	**0.857(12)**	**0.8718(2)**	**0.7078(10)**

Table 5.4: *AUC values for Logistic Regression (LR), Random Forests (RF), Support Vector Machines (SVM) and J48 Decision Trees (J48) classifiers with interest, graph and interest+graph based features using 2:1 spread for the 2,000 user dataset. Here we have assumed that k% links are known in the test set, where k is 10, 25 and 50, respectively. K% known links are used to construct graph features and interest+graph features.*

Exp	Features	Logistic	SVM	RandomForest	J48
4(b)	Interest	0.5666(3)	0.6142(9)	0.5916(12)	0.5844(9)
5(b)	Graph 10%	**0.6748**	**0.7048**	0.662	0.533
6(b)	Graph+Interest 10%	0.6552(3)	0.6638(12)	**0.6722(9)**	**0.5796(3)**
5(b)	Graph 25%	0.764	**0.7712**	0.6608	0.5728
6(b)	Graph+Interest 25%	**0.7768(2)**	0.757(12)	**0.7584(2)**	**0.6208(6)**
5(b)	Graph 50%	0.8808	**0.8674**	0.7926	0.6284
6(b)	Graph+Interest 50%	**0.8886(2)**	0.857(12)	**0.8718(2)**	**0.6706(6)**

known, interest+graph with 10% links known and interest+graph with 50% links known for 2,000 users dataset as compared to 1,000 users dataset. Results obtained from Support Vector Machine with 1:1 and 2:1 spread for 1,000 users dataset outperforms 2,000 users dataset, while, results obtained for 2,000 user dataset from Logistic Regression classifier with 2:1 spread for graph features with 50% links known, interest+graph features with 25% and 50% links known outperformed results obtained for 1,000 user dataset.

Figures 5.3 and 5.4 are the AUC plots for the dataset containing 2,000 users of *LiveJournal* social network for all the four classifiers with 1:1 and 2:1 spread, respectively. Graphs

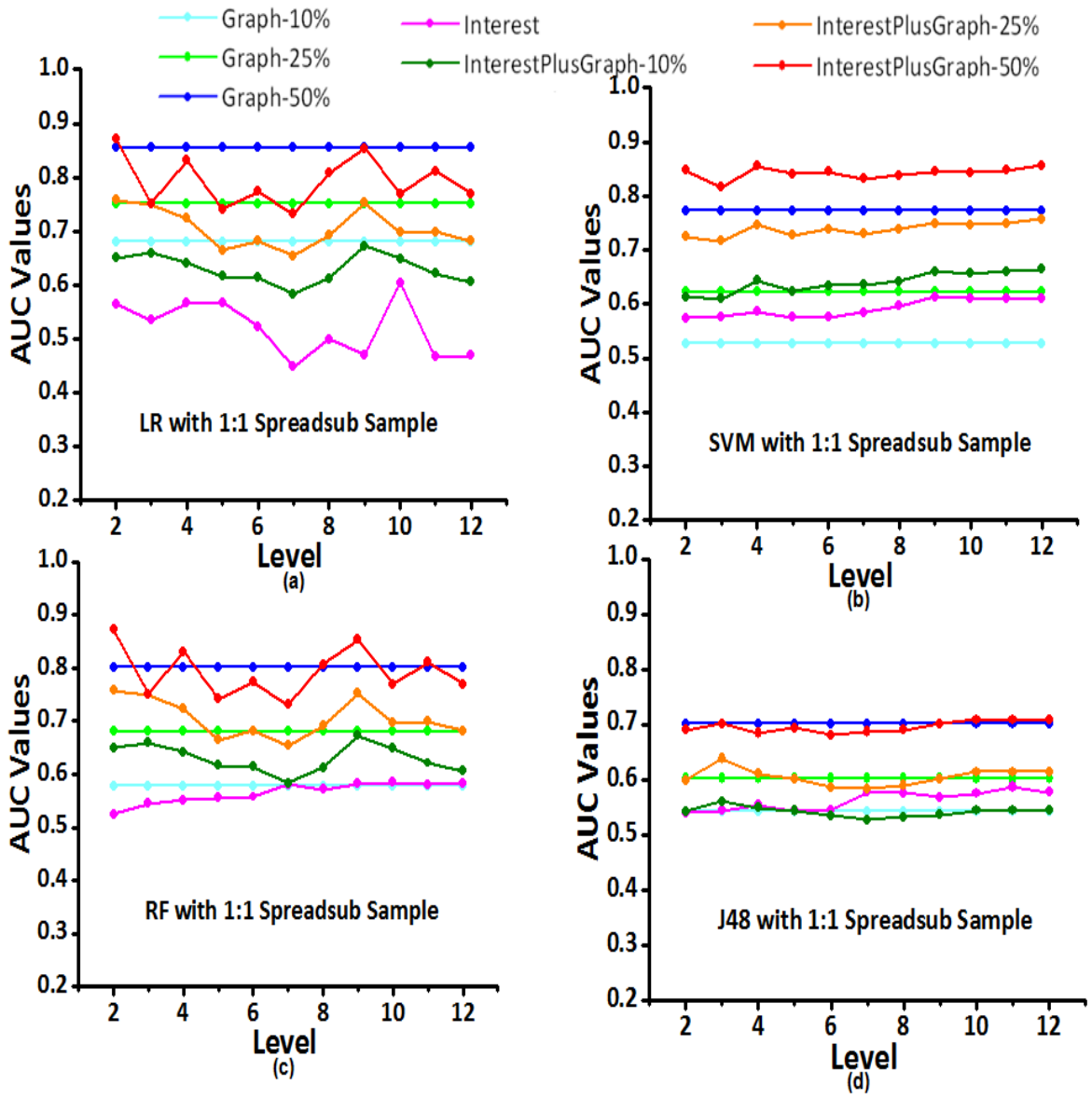

Figure 5.3: *Graphs obtained by plotting AUC values against the number of levels in the interest ontology, (a) Logistic Regression with 1:1 spread, (b) Support Vector Machine classifier with 1:1 spread, (c) Random Forest with 1:1 spread, and (d) J48 classifier with 1:1 spread for 2,000 users dataset.*

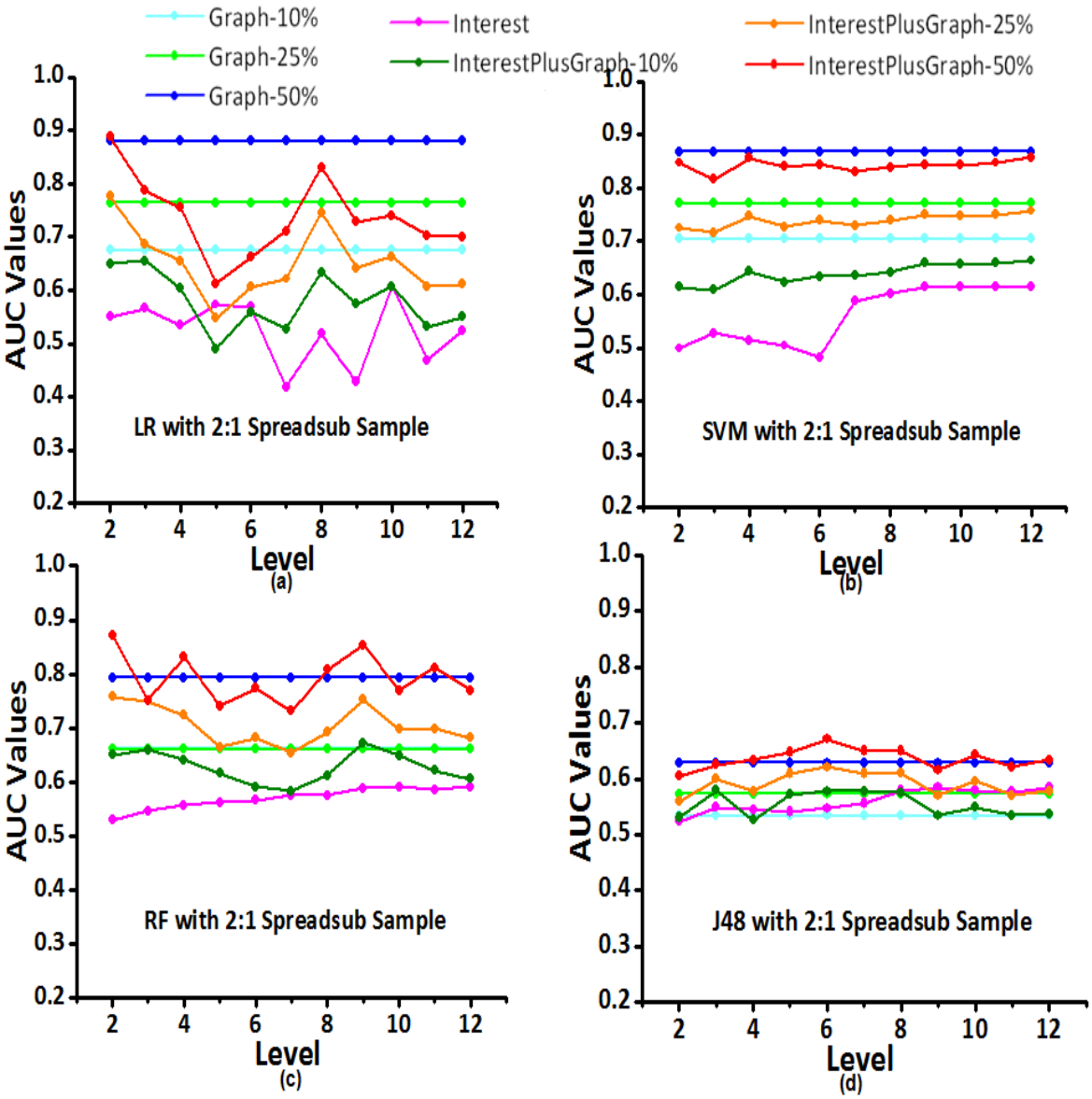

Figure 5.4: *Graphs obtained by plotting AUC values against the number of levels in the interest ontology, (a) Logistic Regression with 2:1 spread, (b) Support Vector Machine classifier with 2:1 spread, (c) Random Forest with 2:1 spread, and (d) J48 classifier with 2:1 spread for 2,000 users dataset.*

are generated by plotting AUC values obtained using graph based features, interest based features and interest+graph based features for 2,000 users of *LiveJournal* social network against different levels in ontology (constructed over interests of 2,000 users). Observing graphs for Logistic Regression classifier with spread 1:1 and 2:1 (in Figure 5.3 (a) and 5.4 (a)) we see that graph features alone outperform interest features. Graph features with 10%, 25%, 50% links outperforms interest+graph features with 10%, 25%, 50% links known for Logistic Regression with 1:1 and 2:1 spread. Although, highest AUC value is observed for interest+graph features with 25% and 50% links known as compared to graph features for Logistic Regression with 1:1 and 2:1 spread but its not consistent across all levels of ontology. Observing graphs for Support Vector Machine with 1:1 spread from Fig 5.3 (b), we see that interest+graph features with 10%, 25% and 50% links outperformed graph features with 10%, 25% and 50% links known. While, graph features with 10%, 25% and 50% links known outperformed all other features for Support Vector Machine classifier with 2:1 spread (Fig. 5.4 (b)). Interest+graph features with 10% links outperformed graph features with 10% link for Random Forest classifier with 1:1 spread (Fig. 5.3 (c)). While, graph feature with 25% and 50% links outperformed interest+graph features for Random Forest classifier with 1:1 spread. The highest AUC value is observed for interest+graph features with 25% and 50% known links, it is not consistent across all levels of the ontology. Graph features with 10%, 25% and 50% links outperformed interest+graph features for Random Forest classifier with 2:1 spread (Fig. 5.4 (c)), although highest AUC value is obtained for interest+graph features but it is not consistent across all ontology levels. Performance pattern for J48 classifier with 2:1 spread is similar to that of Random Forest classifier with 2:1 spread. Furthermore, graph features with 10%, 25% and 50% links outperforms interest+graph features for J48 classifier with 1:1 spread. Although, highest AUC value is observed for interest+graph features with 10%, 25% and 50% links known as compared to graph features but these values are not consistent across all levels of the ontology (Fig. 5.3 (d) and 5.4 (d)).

5.1.3 Results from 3000 user dataset

The AUC values reported by running experiment 7(a), 8(a) and 9(a) using Logistic Regression with 1:1 spread, Support Vector Machine with 1:1 spread, Random Forest with 1:1 spread and J48 Decision Trees with 1:1 spread are reported in Table 5.5. The AUC values reported by running experiment 7(b), 8(b) and 9(b) using Logistic Regression with 2:1 spread, Support Vector Machine with 2:1 spread, Random Forest with 2:1 spread and J48 Decision Trees with 2:1 spread are reported in Table 5.5.

Table 5.5: *AUC values for Logistic Regression (LR), Random Forests (RF), Support Vector Machines (SVM) and J48 Decision Trees (J48) classifiers with interest, graph and interest+graph based features using 1:1 spread for the 3,000 user dataset. Here we have assumed that k% links are known in the test set, where k is 10, 25 and 50, respectively. K% known links are used to construct graph features and interest+graph features.*

Exp	Features	Logistic	SVM	RandomForest	J48
7(a)	Interest	0.561(10)	0.6136(9)	0.6048(12)	0.6212(11)
8(a)	Graph 10%	**0.6722**	0.6414	0.5788	0.5826
9(a)	Graph+Interest 10%	0.6586(2)	**0.6638(12)**	**0.6402(12)**	**0.604(8)**
8(a)	Graph 25%	0.7674	0.7338	0.6962	0.6608
9(a)	Graph+Interest 25%	**0.7714(2)**	**0.757(12)**	**0.7484(3)**	**0.6794(8)**
8(a)	Graph 50%	0.8696	0.8282	0.8154	0.7334
9(a)	Graph+Interest 50%	**0.8742(2)**	**0.857(12)**	**0.8494(3)**	**0.7508(8)**

Table 5.6: *AUC values for Logistic Regression (LR), Random Forests (RF), Support Vector Machines (SVM) and J48 Decision Trees (J48) classifiers with interest, graph and interest+graph based features using 2:1 spread for the 3,000 user dataset. Here we have assumed that k% links are known in the test set, where k is 10, 25 and 50, respectively. K% known links are used to construct graph features and interest+graph features.*

Exp	Features	Logistic	SVM	RandomForest	J48
7(b)	Interest	0.5418(2)	0.6424(9)	0.613(9)	0.6178(10)
8(b)	Graph 10%	**0.7002**	0.6208	**0.662**	**0.5684**
9(b)	Graph+Interest 10%	0.626(10)	**0.6526(9)**	0.624(2)	0.5652(10)
8(b)	Graph 25%	**0.7914**	0.734	0.6826	0.629
9(b)	Graph+Interest 25%	0.7132(9)	**0.7518(9)**	**0.7484(3)**	**0.6514(10)**
8(b)	Graph 50%	**0.8844**	0.853	0.8088	0.6284
9(b)	Graph+Interest 50%	0.8394(9)	**0.8628(9)**	**0.8494(3)**	**0.7304(11)**

Analyzing the AUC values presented in Table 5.5 and 5.6 we can see that interest+graph

based features outperformed interest based features alone. For Logistic Regression with 1:1 spread graph features with 10% known links outperformed interest+graph features with 10% known links. While, interest+graph features with 25% and 50% known links outperformed graph features with 25% and 50% links known. However, for classifiers Support Vector Machine, Random Forest and J48 with spread 1:1 interest+graph features outperformed graph features. Observing Table 5.6, we see that graph based features outperformed interest+graph features for Logistic Regression classifier. While, interest+graph features outperformed graph features for Support Vector Machine classifier with 2:1 spread. Furthermore, graph features with 10% links known outperformed interest+graph with 10% links known and interest+graph features with 25% and 50% links known outperformed graph features for Random Forest and J48 classifiers.

Figures 5.5 and 5.6 are the AUC plots for the dataset containing 3,000 users of *LiveJournal* social network for all the four classifiers with 1:1 and 2:1 spread. Observing these graphs we see that for Logistic Regression classifier with 1:1 and 2:1 spread (Fig. 5.5 (a) and 5.6 (a)) graph based features with 10%, 25% and 50% links known outperformed interest+graph based features with 10%, 25% and 50% links known. Although, highest AUC value is reported by interest+graph based features, this value is not consistent across all levels of the ontology. For Support Vector Machine with spread 1:1 and 2:1 interest+graph features with 50% links known outperformed graph features with 50% links known. However, graph features with 25% and 50% known links outperformed interest+graph features with 25% and 50% known links. Although, highest AUC value is obtained for interest+graph features it is not consistent across all levels of the ontology (except levels 9 through 12). For Random Forest classifier with 1:1 spread (Fig. 5.5 (c) and 5.6 (c)) interest+graph based features outperformed graph features alone and interest features alone with consistent AUC values across all levels of the ontology. While, with 2:1 spread interest+graph based features with 25% and 50% links known outperformed graph based features with consistent AUC values across all levels of the ontology. Furthermore, for J48 classifier with 1:1 and 2:1 spread

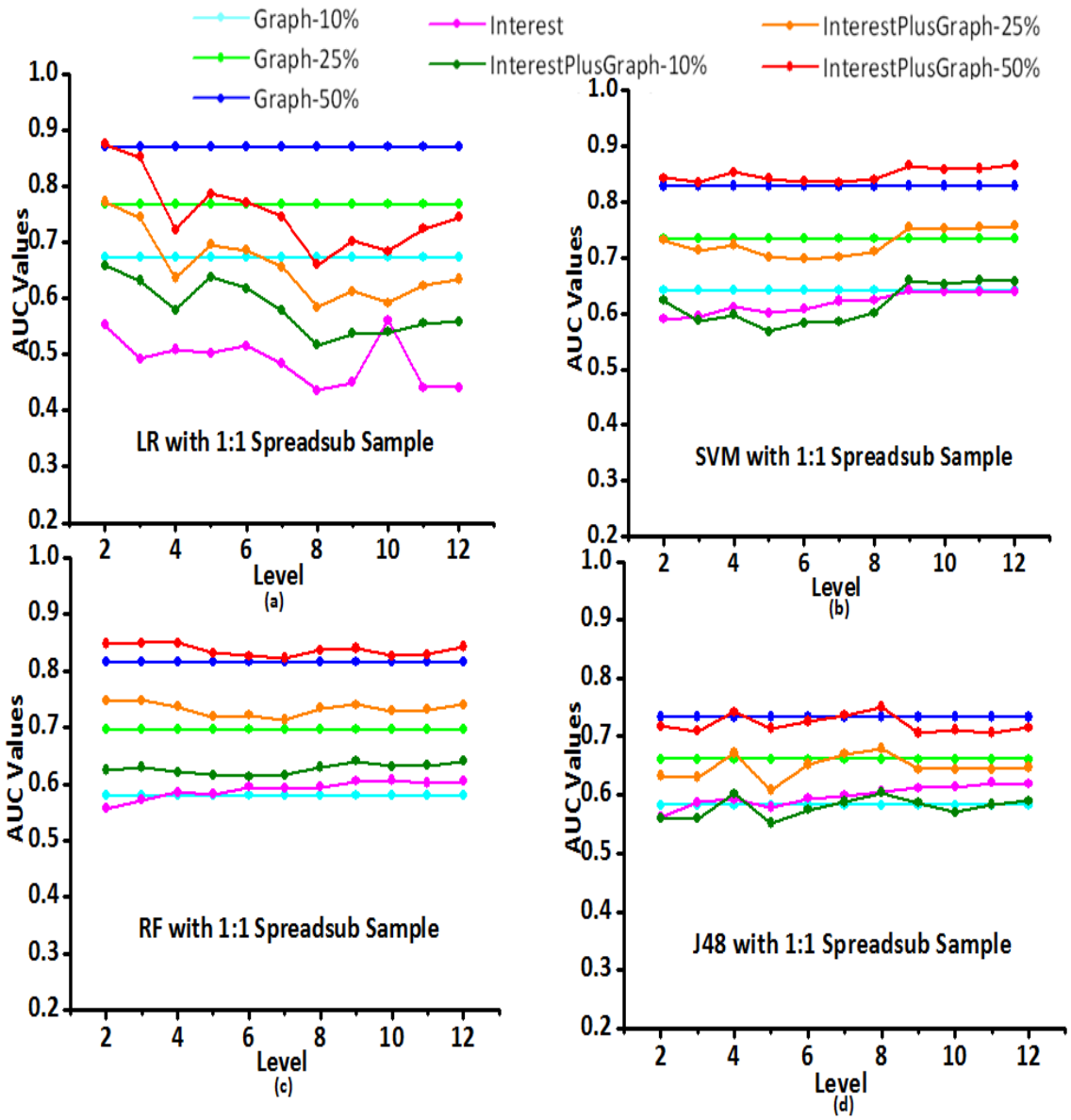

Figure 5.5: *Graphs obtained by plotting AUC values against the number of levels in the interest ontology, (a) Logistic Regression with 1:1 spread, (b) Support Vector Machine classifier with 1:1 spread, (c) Random Forest with 1:1 spread, and (d) J48 classifier with 1:1 spread for 3,000 users dataset.*

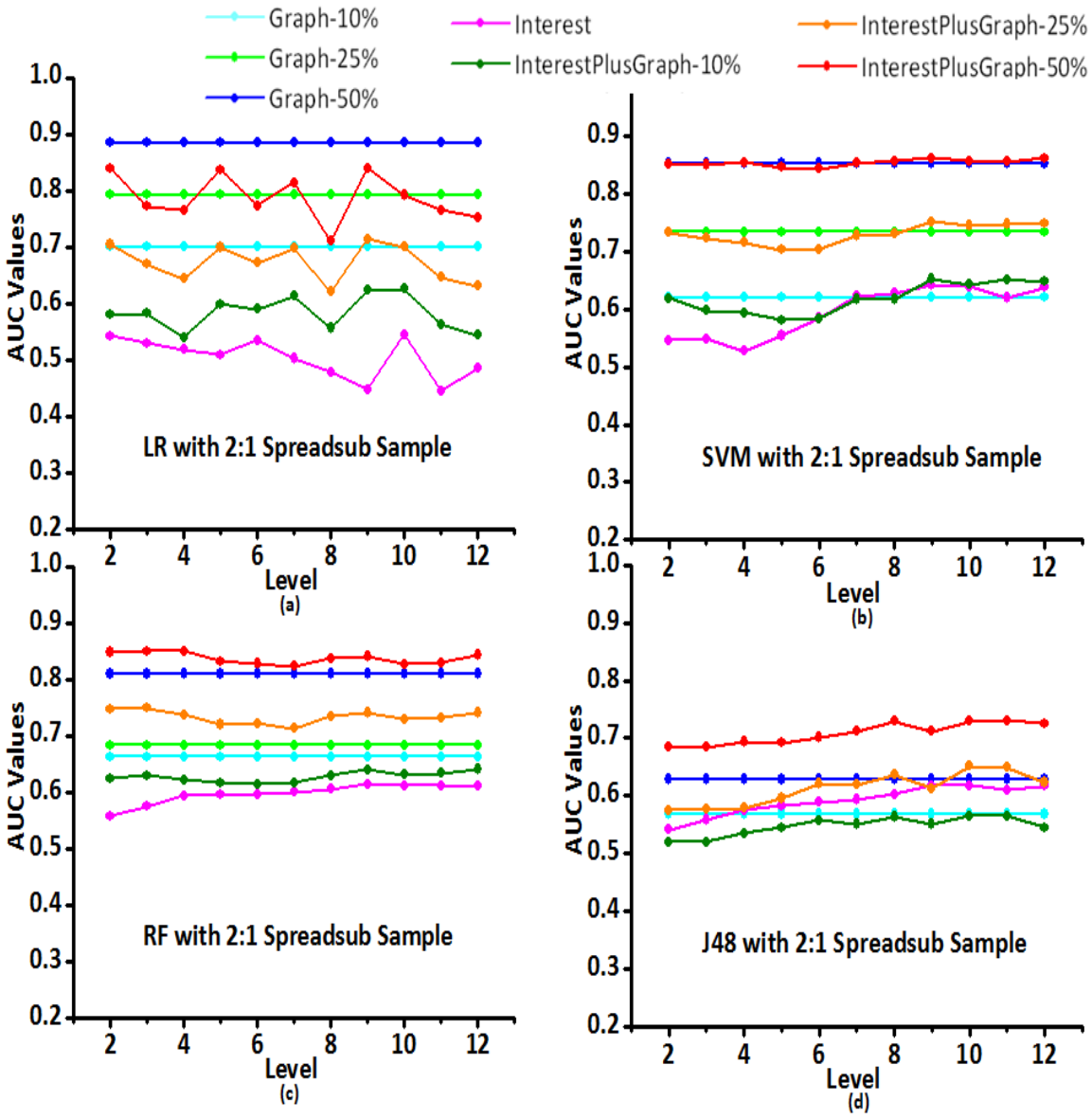

Figure 5.6: *Graphs obtained by plotting AUC values against the number of levels in the interest ontology, (a) Logistic Regression with 2:1 spread, (b) Support Vector Machine classifier with 2:1 spread, (c) Random Forest with 2:1 spread, and (d) J48 classifier with 2:1 spread for 3,000 users dataset.*

showed performance similar to Random Forest.

5.1.4 Results from 4000 user dataset

The AUC values reported by running experiment 10(a), 11(a) and 12(a) using Logistic Regression with 1:1 spread, Support Vector Machine with 1:1 spread, Random Forest with 1:1 spread and J48 Decision Trees with 1:1 spread are reported in Table 5.7. In the above table, we compare the AUC values obtained by using interest based features, graph based features and interest+graph based features. For 4000 user dataset, we did not perform experiments using classifiers with 2:1 spread, because the classifiers do not show much improvement in the performance and often show poor behavior.

Table 5.7: *AUC values for Logistic Regression (LR), Random Forests (RF), Support Vector Machines (SVM) and J48 Decision Trees (J48) classifiers with interest, graph and interest+graph based features using 1:1 spread for the 4,000 user dataset. Here we have assumed that k% links are known in the test set, where k is 10, 25 and 50, respectively. K% known links are used to construct graph features and interest+graph features.*

Exp	Features	Logistic	SVM	RandomForest	J48
10(a)	Interest	0.578(12)	0.6508(12)	0.6172(12)	0.624(9)
11(a)	Graph 10%	0.6974	0.5204	0.587	0.6474
12(a)	Graph+Interest 10%	**0.7476(2)**	**0.7388(11)**	**0.622(11)**	**0.6474(2)**
11(a)	Graph 25%	0.791	0.6568	0.7146	**0.715**
12(a)	Graph+Interest 25%	**0.81588(2)**	**0.8352(11)**	**0.7572(5)**	0.7136(5)
11(a)	Graph 50%	0.8922	0.8152	0.8278	0.7664
12(a)	Graph+Interest 50%	**0.8922(2)**	**0.9116(9)**	**0.8512(2)**	**0.7686(11)**

Analyzing the Table 5.7 we observe that, interet+graph features with 10%, 25% and 50% links known outperformed graph based features with 10%, 25% and 50% and interest based features for Logistic Regression, Support Vector Machine and Random Forest classifiers with 1:1 spread. While, for J48 classifier with 1:1 spread interest+graph features with 10% and 50% known links outperformed graph based features with 10% and 50% links known and graph features with 25% links known outperformed interest+graph features with 25% links known.

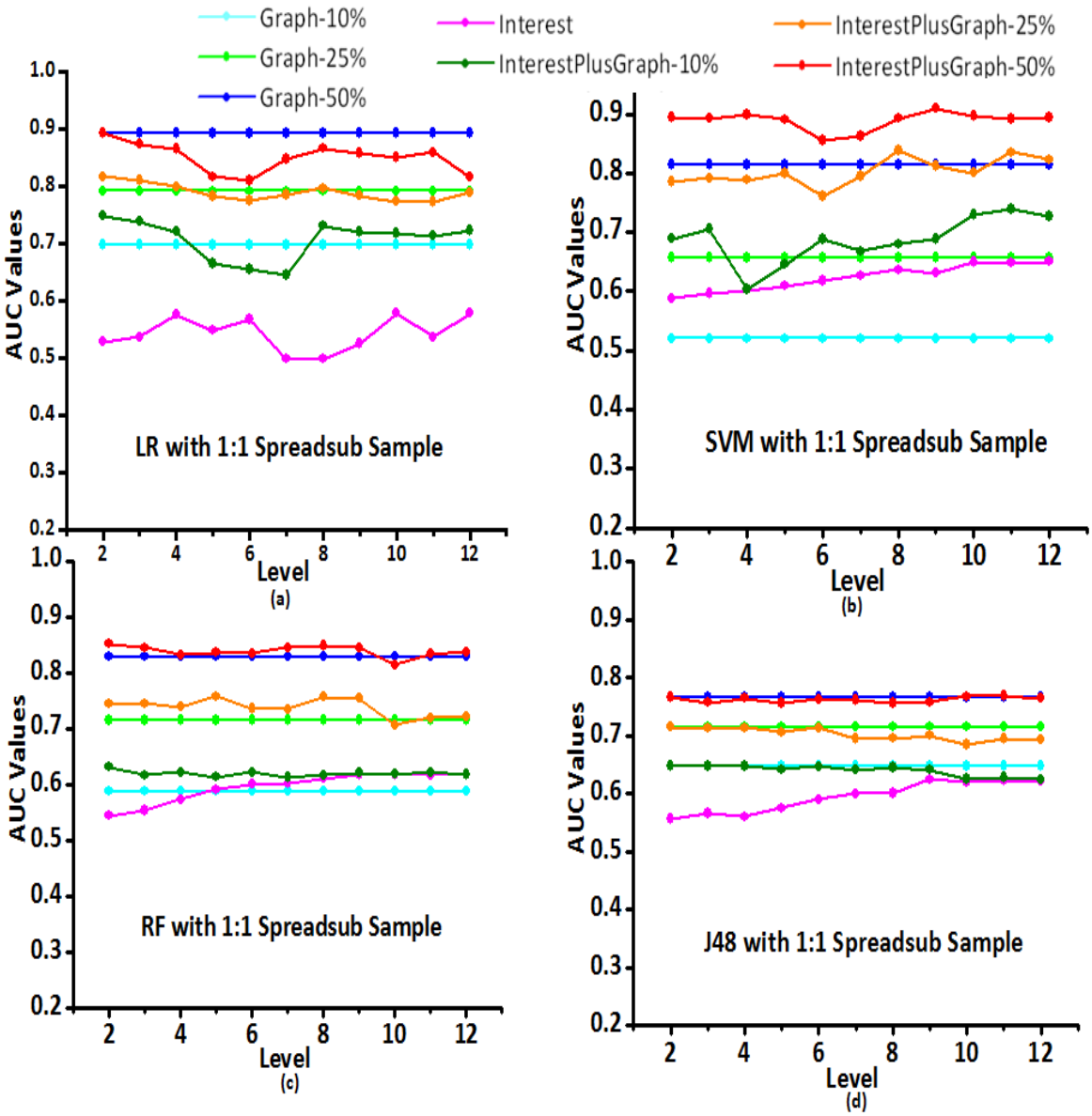

Figure 5.7: *Graphs obtained by plotting AUC values against the number of levels in the interest ontology, (a) Logistic Regression with 1:1 spread, (b) Support Vector Machine classifier with 1:1 spread, (c) Random Forest with 1:1 spread, and (d) J48 classifier with 1:1 spread for 4000 users dataset.*

Figures 5.7 shows the AUC plots for dataset containing 4,000 users of the *LiveJournal* social network for all four classifiers with 1:1 spread. Observing the graph we see that, for Logistic Regression with 1:1 (Fig. 5.7 (a)) interest+graph based features with 10%, 25% and 50% links known outperformed graph based features with 10%, 25% and 50% links known. Although, highest AUC value is reported by interest+graph based features, this value is not consistent across all levels of the ontology. Observing results obtained for Support Vector Machine classifier with 1:1 spread (in Fig. 5.7 (b)) suggests that interest+graph features with 10%, 25% and 50% known links outperforms interest features and graph features with 10%, 25% and 50% links known with consistent higher values at all levels of the ontology. Similar observation was shown by Random Forest classifier with 1:1 spread (Fig. 5.7 (c)). Although, highest AUC value is shown by interest+graph based features, this behavior is not consistent across all levels of the ontology. Looking at the graph reported by J48 classifier with 1:1 spread in Figure 5.7, we see that interest+graph features with 10% and 50% known links outperformed graph based features with 10% and 50% links known and graph features with 25% links known outperformed interest+graph features with 25% links known, but the AUC values are not consistent across all ontology levels.

5.2 Study of the Effect of Larger Datasets on the Performance

We analyze the previous results in terms of performance for varying the size of datasets. Generation of these data sets was already seen in Chapter 4 Section 4.1. One of the main goals of this thesis is to study the performance of classifiers at the task of predicting friendship links when presented with graph based features, interest based features and interest+graph based features generated over all the data sets. We have organized this section in three parts: in Section 5.2.1 we compare the performance of all classifiers over graph based features for all four data sets; in Section 5.2.2 we compare the performance of all classifiers over interest based features for all four data sets; and in Section 5.2.3 we compare the performance of all

classifiers generally over interest+graph based features as the number of users increases.

5.2.1 The Effect of Graph Based Features on the Performance

Comparison of the performance of graph based features with 10%, 25% and 50% known links over 1000, 2000, 3000 and 4000 users data sets of *LiveJournal* social network using Logistic Regression, Support Vector Machine, Random Forest and J48 classifiers with 1:1 spread is presented in Table 5.8, 5.9, 5.10 and 5.11, respectively.

Table 5.8: *Comparing AUC values obtained from Logistic Regression classifier with 1:1 spread for graph based features with 10%, 25%, and 50% links known for 1000, 2000, 3000 and 4000 users data sets.*

Features	1000	2000	3000	4000
Graph 10%	**0.7258**	0.6798	0.6722	0.6974
Graph 25%	0.7624	0.7508	0.7674	**0.791**
Graph 50%	0.8538	0.855	0.8696	**0.8922**

Table 5.9: *Comparing AUC values obtained from Support Vector Machine classifier with 1:1 spread using graph based features with 10%, 25%, and 50% links known for 1000, 2000, 3000 and 4000 users data sets.*

Features	1000	2000	3000	4000
Graph 10%	**0.7418**	0.5264	0.6414	0.5204
Graph 25%	**0.7924**	0.622	0.7338	0.6568
Graph 50%	**0.8624**	0.7736	0.8282	0.8152

Table 5.10: *Comparing AUC values obtained from Random Forest classifier with 1:1 spread using graph based features with 10%, 25%, and 50% links known for 1000, 2000, 3000 and 4000 users data sets.*

Features	1000	2000	3000	4000
Graph 10%	0.5698	0.5786	0.5788	**0.587**
Graph 25%	0.7036	0.681	0.6962	**0.7146**
Graph 50%	0.7798	0.8012	0.8154	**0.8278**

Observing these tables we see that, performance for all classifier does not improves as the number of users increases in the data set. Results from Support Vector Machine

Table 5.11: *Comparing AUC values obtained from J48 classifier with 1:1 spread using graph based features with 10%, 25%, and 50% links known for 1000, 2000, 3000 and 4000 users data sets.*

Features	1000	2000	3000	4000
Graph 10%	0.5102	0.5414	0.5826	**0.6474**
Graph 25%	0.4838	0.6018	0.6608	**0.715**
Graph 50%	0.5254	0.7008	0.7334	**0.7664**

showed degradation when number of users were increased in the data sets. While, Logistic Regression showed degradation in the performance for graph features with 10% links known but an improvement for graph features with 25% and 50% links known. However, Random Forest and J48 classifiers showed an improvement in the performance for 4000 user data set.

5.2.2 The Effect of Interest Based Features on the Performance

In this section we compare and examine the performance of classifiers using interest based features generated over 1000, 2000, 3000 and 4000 user data sets. Table 5.12 compares results obtained for 1000, 2000, 3000 and 4000 data sets from Logistic Regression classifier with 1:1 spread using interest based features. Studying this table we can see that the classifier does not show a drastic increase in its performance when the number of users increases. Highest AUC value is shown by 2000 user data set. We plot graph (Fig. 5.8) obtained by reporting AUC values against the number of levels in the interest ontology so as to compare the results obtained for Logistic Regression classifier with 1:1 spread over 1000, 2000, 3000 and 4000 user data sets. The graph depicts a very in-consistent performance by the classifier.

Table 5.12: *Comparing AUC values obtained from Logistic Regression classifier with 1:1 spread using interest based features for 1000, 2000, 3000 and 4000 user data sets.*

Features	1000	2000	3000	4000
Interest	0.564(4)	**0.6042(10)**	0.561(10)	0.578(12)

Table 5.13 compares results obtained for 1000, 2000, 3000 and 4000 data sets from Support Vector Machine classifier with 1:1 spread using interest based features. Studying this table, we can see that the classifier shows a gradual increase in the performance when

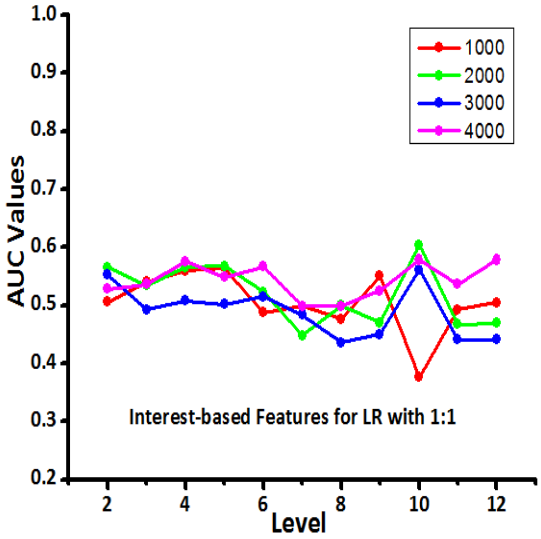

Figure 5.8: *Comparing graphs obtained by reporting AUC values against the number of levels in interest ontology for Logistic Regression classifier with 1:1 spread using interest based features over 1000, 2000, 3000 and 4000 user data sets.*

the number of users in the data set increases. We also plot graph (Fig. 5.9) obtained by reporting AUC values against the number of levels in the interest ontology so as to compare the results obtained for Support Vector Machine classifier with 1:1 spread over 1000, 2000, 3000 and 4000 user data sets. The graph depicts a consistent increase in the performance of the classifier with increase in the size of data sets, 3000 and 4000 user data sets outperformed 1000 and 2000 user data sets. Although, highest AUC value is reported by 4000 user data set, it is not consistent across all levels of the ontology.

Table 5.13: *Comparing AUC values obtained from Support Vector Machine classifier with 1:1 spread using interest based features for 1000, 2000, 3000 and 4000 user data sets.*

Features	1000	2000	3000	4000
Interest	0.5718(11)	0.6136(9)	0.6136(9)	**0.6508(12)**

We compare results obtained for 1000, 2000, 3000 and 4000 data sets from Random Forest classifier with 1:1 spread using interest based features in Table 5.14. Observing this table we can see that these AUC values increase when the number of users in the data set

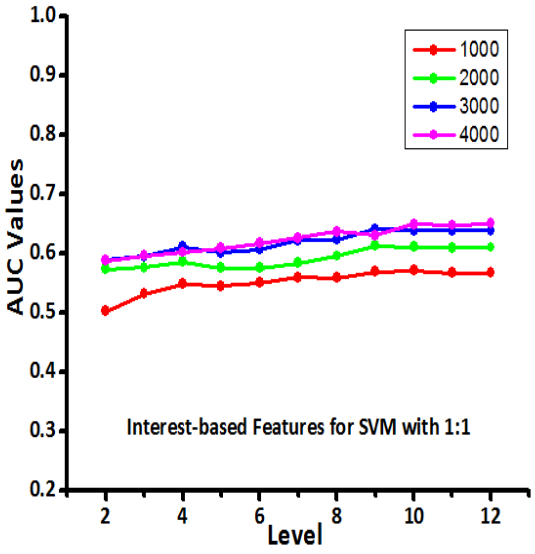

Figure 5.9: *Comparing graphs obtained by reporting AUC values against the number of levels in interest ontology for Support Vector Machine classifier with 1:1 spread using interest based features over 1000, 2000, 3000 and 4000 user data sets.*

also increases. We also plot a graph obtained by reporting AUC values against the number of levels in the interest ontology so as to compare the results obtained for Random Forest classifier with 1:1 spread over 1000, 2000, 3000 and 4000 user data sets. The graph obtained in Figure 5.10 shows that the AUC values obtained for 4000 user data set outperforms all other data sets.

Table 5.14: *Comparing AUC values obtained from Random Forest classifier with 1:1 spread using interest based features for 1000, 2000, 3000 and 4000 user data sets.*

Features	1000	2000	3000	4000
Interest	0.5536(11)	0.5848(10)	0.6048(12)	**0.6172(12)**

Table 5.15: *Comparing AUC values obtained from J48 Decision Tree classifier with 1:1 spread using interest based features for 1000, 2000, 3000 and 4000 user data sets.*

Features	1000	2000	3000	4000
Interest	0.5454(12)	0.5862(11)	0.6212(11)	**0.6226(11)**

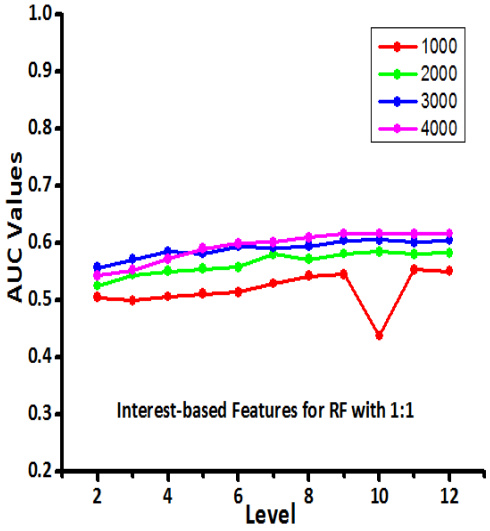

Figure 5.10: *Comparing graphs obtained by reporting AUC values against the number of levels in interest ontology for Random Forest classifier with 1:1 spread using interest based features over 1000, 2000, 3000 and 4000 user data sets.*

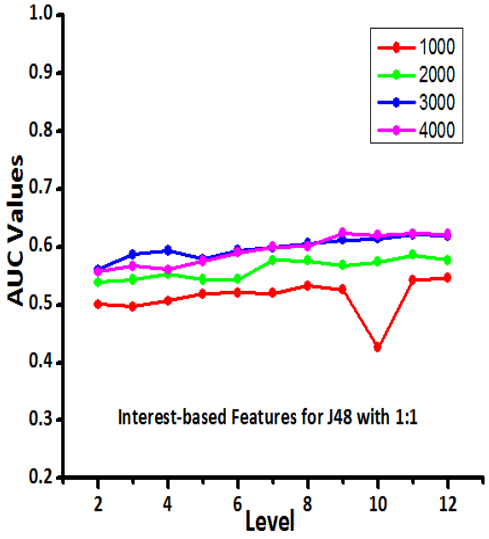

Figure 5.11: *Comparing graphs obtained by reporting AUC values against the number of levels in interest ontology for J48 Decision Trees classifier with 1:1 spread using interest based features over 1000, 2000, 3000 and 4000 user data sets.*

We study Table 5.15, which compares results obtained for 1000, 2000, 3000 and 4000 data sets from J48 Decision Tree classifier with 1:1 spread using interest based features. We see that the performance of the classifier improves as we increase the number of users in the data set with highest AUC value obtained for 4000 user data set. Furthermore, we also plot graph (Fig. 5.11) obtained by reporting AUC values against the number of levels in the interest ontology so as to compare the results obtained for J48 classifier with 1:1 spread over 1000, 2000, 3000 and 4000 user data sets. The graph depicts a consistent increase in the performance of the classifier with increase in the size of data sets, 3000 and 4000 user data sets outperformed 1000 and 2000 user data sets. Although, highest AUC value was reported by 4000 user data set, it is not consistent across all levels of the ontology.

5.2.3 The Effect of Graph & Interest Based Features on the Performance

In this section we compare the results obtained by using interest+graph based features from all data sets for each classifier. We compare results from 1000, 2000, 3000 and 4000 data sets for Logistic Regression classifier with 1:1 spread using interest+graph based features generated over these data sets. Similarly, we compare results obtained for Support Vector Machine, Random Forest and J48 Decision Tree classifiers with 1:1 spread.

Table 5.16: *Comparing AUC values obtained from Logistic Regression classifier with 1:1 spread using interest+graph based features with 10%, 25% and 50% links known for 1000, 2000, 3000 and 4000 user data sets.*

Features	1000	2000	3000	4000
Interest & Graph 10%	0.6744(8)	0.6722(9)	0.6786(4)	**0.7476(2)**
Interest & Graph 25%	0.78(8)	0.7584(2)	0.7714(2)	**0.81588(2)**
Interest & Graph 50%	0.8878(8)	0.8718(2)	0.8742(2)	**0.8922(2)**

Looking at the comparison made in Table 5.16, we can infer that the highest AUC value is obtained for 4000 data set using Logistic Regression classifier with 1:1 spread. From graph (Fig. 5.12) obtained by reporting AUC values against the number of levels in the interest ontology so as to compare the results obtained for Logistic Regression classifier with 1:1

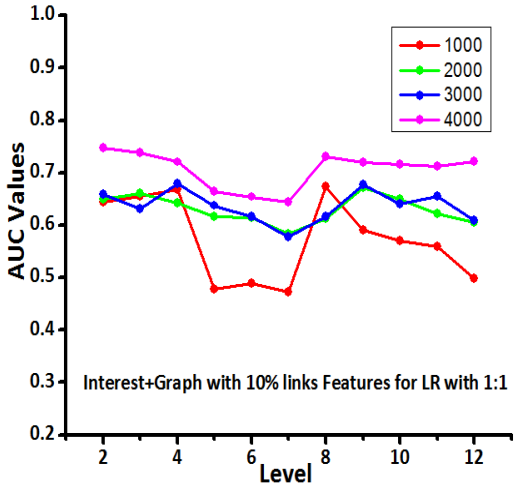

Figure 5.12: *Comparing graphs obtained by reporting AUC values against the number of levels in interest ontology for Logistic Regression classifier with 1:1 spread using interest+graph based features with 10% links known over 1000, 2000, 3000 and 4000 user data sets.*

spread using interest+graph features with 10% links known over 1000, 2000, 3000 and 4000 user data sets, we can see that AUC values obtained for 4000 user data set outperformed all other data sets. A similar observation was made for interest+graph features with 25% and 50% links known from Table 5.16 and 5.16, and Figure 5.13 and 5.14, respectively. Although, the highest AUC value was obtained for Logistic Regression using interest+graph based features for 4000 user data set, it is not consistent across all ontology levels.

Table 5.17: *Comparing AUC values obtained from Support Vector Machine classifier with 1:1 spread using interest+graph based features with 10%, 25% and 50% links known for 1000, 2000, 3000 and 4000 user data sets.*

Features	1000	2000	3000	4000
Interest & Graph 10%	0.7376(12)	0.6638(12)	0.6578(12)	**0.7388(11)**
Interest & Graph 25%	0.8258(12)	0.757(12)	0.757(12)	**0.8382(8)**
Interest & Graph 50%	0.905(10)	0.857(12)	0.8658(12)	**0.9116(9)**

Results obtained from Support Vector Machine with 1:1 spread using interest+graph

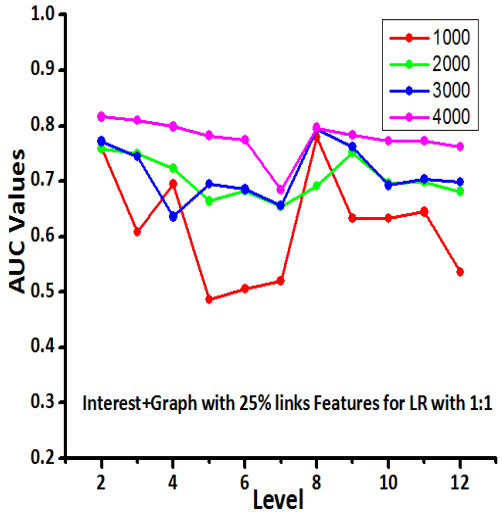

Figure 5.13: *Comparing graphs obtained by reporting AUC values against the number of levels in interest ontology for Logistic Regression classifier with 1:1 spread using interest+graph based features with 25% links known over 1000, 2000, 3000 and 4000 user data sets.*

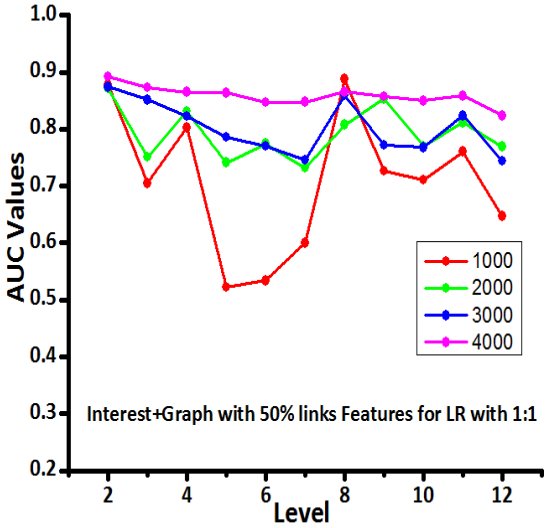

Figure 5.14: *Comparing graphs obtained by reporting AUC values against the number of levels in interest ontology for Logistic Regression classifier with 1:1 spread using interest+graph based features with 50% links known over 1000, 2000, 3000 and 4000 user data sets.*

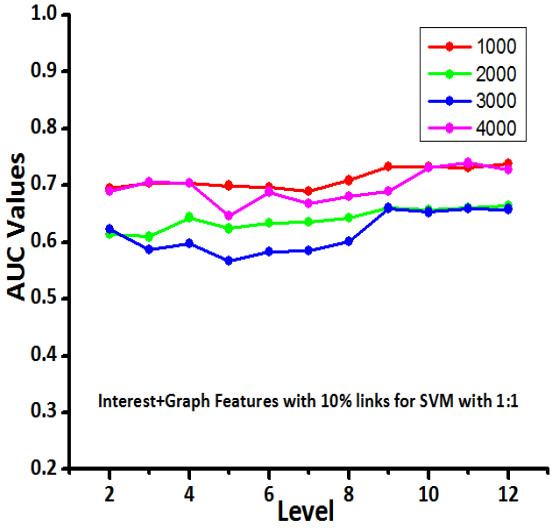

Figure 5.15: *Comparing graphs obtained by reporting AUC values against the number of levels in interest ontology for Support Vector Machine classifier with 1:1 spread using interest+graph based features with 10% links known over 1000, 2000, 3000 and 4000 user data sets.*

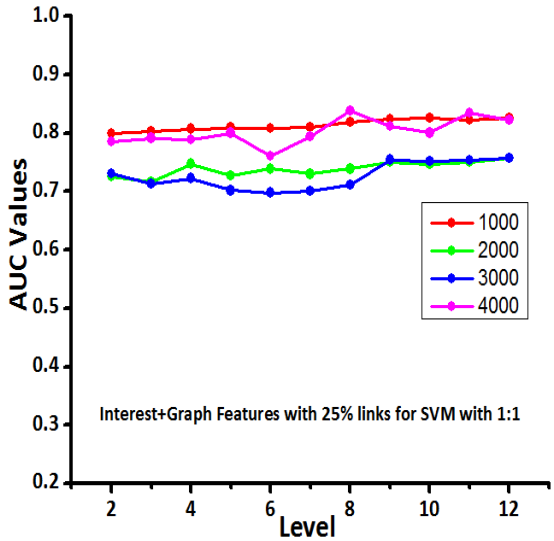

Figure 5.16: *Comparing graphs obtained by reporting AUC values against the number of levels in interest ontology for Support Vector Machine classifier with 1:1 spread using interest+graph based features with 25% links known over 1000, 2000, 3000 and 4000 user data sets.*

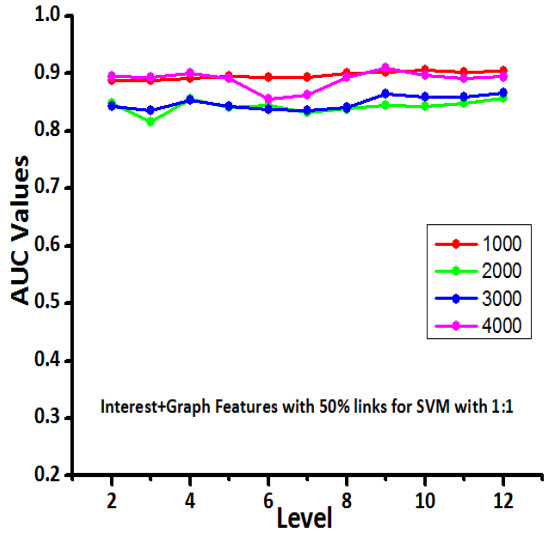

Figure 5.17: *Comparing graphs obtained by reporting AUC values against the number of levels in interest ontology for Support Vector Machine classifier with 1:1 spread using interest+graph based features with 50% links known over 1000, 2000, 3000 and 4000 user data sets.*

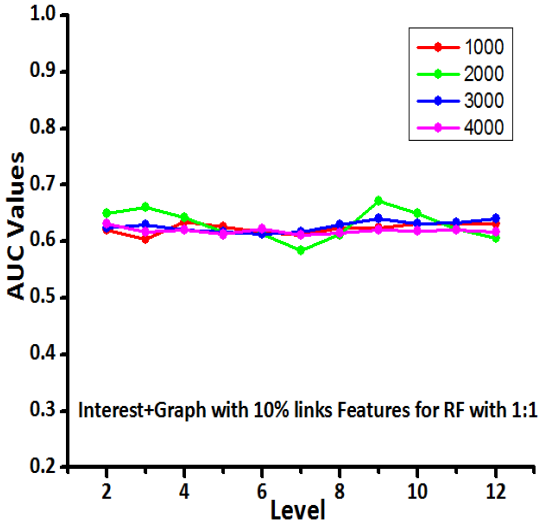

Figure 5.18: *Comparing graphs obtained by reporting AUC values against the number of levels in interest ontology for Random Forest classifier with 1:1 spread using interest+graph based features with 10% links known over 1000, 2000, 3000 and 4000 user data sets.*

features with 10%, 25% and 50% links known over 1000, 2000, 3000 and 4000 user data sets are presented in Table 5.17. These tables suggests that 4000 user data set outperformed 1000, 2000 and 3000 user data sets, as can be observed from the graphs plotted in Figures 5.15, 5.16 and 5.17. Although, the highest AUC value is obtained for features generated for 4000 user data set, it is not consistent across all ontology levels.

Table 5.18: *Comparing AUC values obtained from Random Forest classifier with 1:1 spread using interest+graph based features with 10%, 25% and 50% links known for 1000, 2000, 3000 and 4000 user data sets.*

Features	1000	2000	3000	4000
Interest & Graph 10%	0.6314(10)	**0.6722(9)**	0.6402(12)	0.6304(2)
Interest & Graph 25%	**0.7926(11)**	0.7584(2)	0.7484(3)	0.7572(4)
Interest & Graph 50%	0.86(9)	**0.8718(2)**	0.8494(3)	0.8512(2)

Table 5.19: *Comparing AUC values obtained from J48 Decision Tree classifier with 1:1 spread using interest+graph based features with 10%, 25% and 50% links known for 1000, 2000, 3000 and 4000 user data sets.*

Features	1000	2000	3000	4000
Interest & Graph 10%	0.5338(2)	0.5604(3)	0.5896(12)	**0.6468(6)**
Interest & Graph 25%	0.4988(4)	0.6374(3)	0.6794(8)	**0.7136(6)**
Interest & Graph 50%	0.521(8)	0.7078(10)	0.7508(8)	**0.7686(11)**

Results from Random Forest classifier as shown in Table 5.18 it can be observed that highest AUC value is obtained for 2000 user data set and for 5.18 highest AUC value is also obtained for 2000 user data set. Similarly, observing the graphs plotted in Figure 5.18, 5.19 and 5.20 we can see that highest AUC value is obtained by 2000 user data set in Fig. 5.18 and fig. 5.20, but it is not consistent across all ontology levels, while Figure 5.19 depicts that results obtained from 1000 user data set outperforms all other data sets.

Table 5.19 compares results obtained from 1000, 2000, 3000 and 4000 user data sets using J48 Decision Tree classifier with 1:1 spread using interest+graph features with 10%, 25% and 50% links known, respectively. The above tables suggests that performance of the classifier improves as the number of users in the data set increases. Thus, out of the

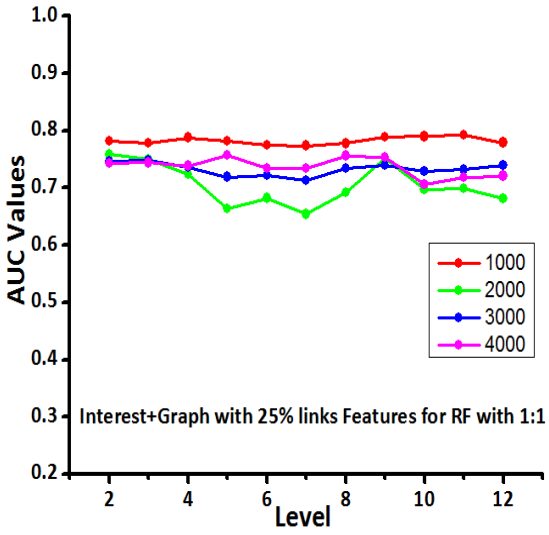

Figure 5.19: *Comparing graphs obtained by reporting AUC values against the number of levels in interest ontology for Random Forest classifier with 1:1 spread using interest+graph based features with 25% links known over 1000, 2000, 3000 and 4000 user data sets.*

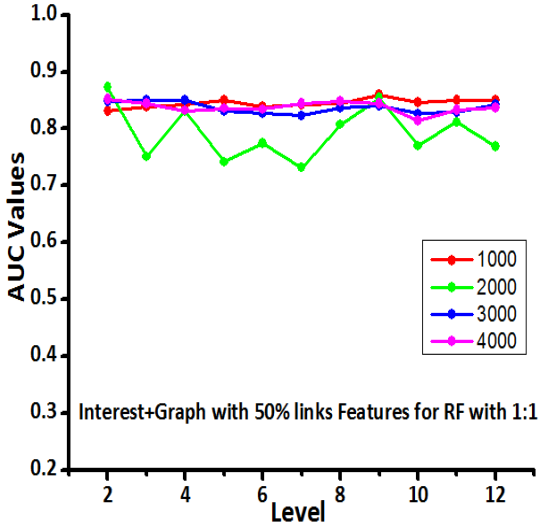

Figure 5.20: *Comparing graphs obtained by reporting AUC values against the number of levels in interest ontology for Random Forest classifier with 1:1 spread using interest+graph based features with 50% links known over 1000, 2000, 3000 and 4000 user data sets.*

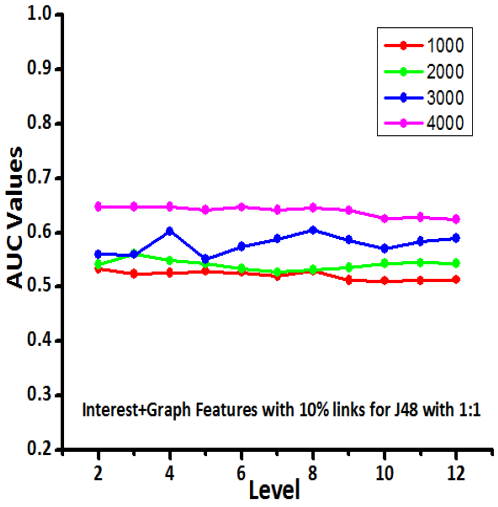

Figure 5.21: *Comparing graphs obtained by reporting AUC values against the number of levels in interest ontology for J48 Decision Tree classifier with 1:1 spread using interest+graph based features with 10% links known over 1000, 2000, 3000 and 4000 user data sets.*

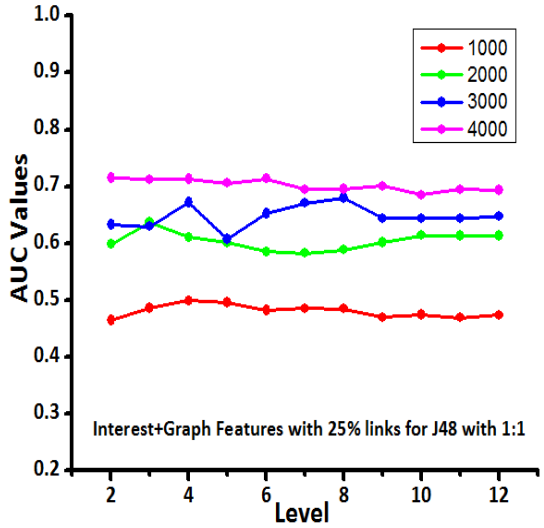

Figure 5.22: *Comparing graphs obtained by reporting AUC values against the number of levels in interest ontology for J48 Decision Tree classifier with 1:1 spread using interest+graph based features with 25% links known over 1000, 2000, 3000 and 4000 user data sets.*

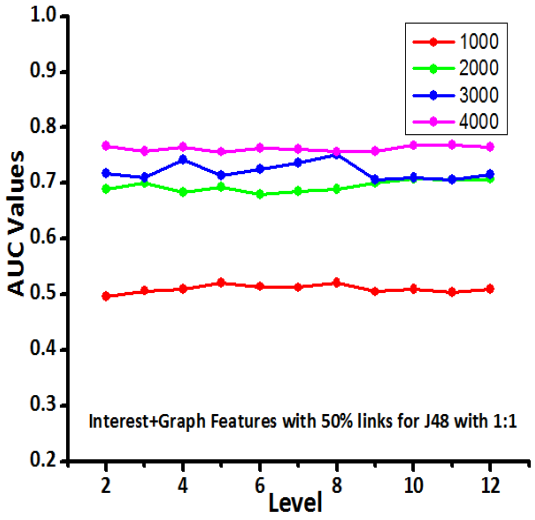

Figure 5.23: *Comparing graphs obtained by reporting AUC values against the number of levels in interest ontology for J48 Decision Tree classifier with 1:1 spread using interest+graph based features with 50% links known over 1000, 2000, 3000 and 4000 user data sets.*

four data sets, 4000 user data set outperforms all other data sets by reporting highest AUC values. Graph plotted in Figure 5.21 suggests that, 4000 user data set outperforms 1000, 2000 and 3000 user data sets with consistent higher AUC values at all levels of the ontology. A similar observation was shown by J48 classifier using interest+graph features with 25% and 50% links known in graphs plotted in Figure 5.22 and 5.23, respectively.

5.3 Comparison with LDA Based Approach

Under this section we compare results generated for 1000 user data set using the ontology approach with the results obtained in the work presented by Parimi [2010]. Parimi [2010] explored the ability of topic modeling techniques such as Latent Dirichlet Allocation (LDA) in order to uncover latent structure in user interests of users of the *LiveJournal* social network. Each user in the dataset was seen as a document and the content of a document corresponded to his/her interests. Thus, each document was treated as a mixture of topics

and each topic in turn was treated as a mixture of words. By using LDA to group interests, Parimi [2010] does not construct an ontology explicitly, but implicitly simulates an ontology by varying the number of latent topics to be identified.

Since, the LDA topic modeling approach presented by Parimi [2010] does not construct an ontology as we have done in this work, we compare the two approaches in this section to study the performance of both the approaches at the task of link prediction in *LiveJournal* social network. We have followed the same procedure to pre-process data before constructing data sets as mention in [Parimi, 2010], these steps are explained in Chapter 4 Section 4.1.

Table 5.20: *Comparing highest AUC values reported by Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM) classifiers with 1:1 spread using interest based features with the interest ontology constructed in our work and using Interest based features generated from LDA topic modeling approach presented by Parimi [2010] over 1,000 users data set.*

Classifiers	Ontology Approach	LDA Topic Modeling
LR	0.564(4)	**0.6258(160)**
RF	0.5536(11)	**0.5808(90)**
SVM	0.5718(11)	**0.6158(90)**

Table 5.20 suggests that interest based features generated from LDA approach give better performance over the classifiers as compared to interest based features generated by using interest ontology. Graphs are generated by plotting AUC values against the ontology levels, and by plotting AUC values against topics for LDA approach for Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM) classifiers with 1:1 spread as shown in Figure 5.24. These graphs indicate that LDA approach performed better than ontology approach when used to generate interest based features alone.

In Table 5.21 and Figure 5.25 we compare the AUC values reported by the Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM) classifiers with 1:1 spread using interest+graph based features with 10% links known generated using ontology approach and LDA topic modeling technique. The table suggests that, AUC values obtained using ontology approach outperformed AUC values obtained using LDA approach. Graphs

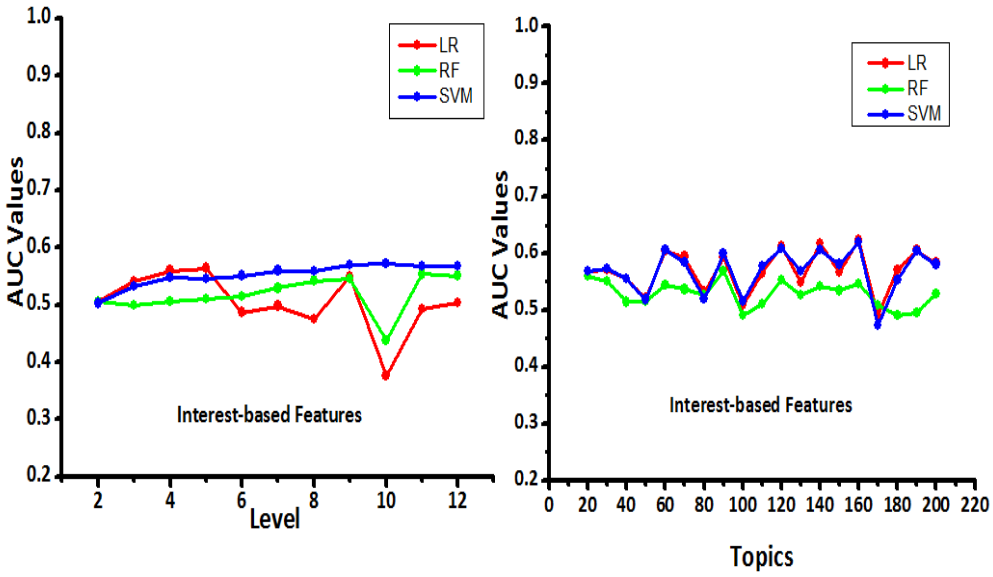

Figure 5.24: *Graphs obtained by reporting AUC values against the number of levels in interest ontology and AUC values against topics from LDA approach for Logistic Regression, (b) Random Forest, and (c) Support Vector Machine classifier with 1:1 spread using interest based features generated for 1,000 users data set.*

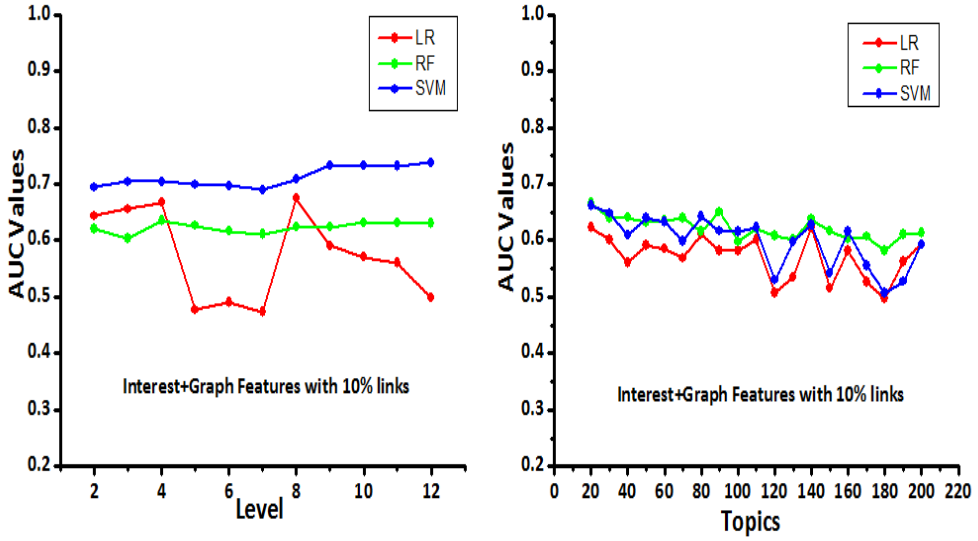

Figure 5.25: *Graph obtained for ontology approach by reporting AUC values against the number of levels and for LDA approach by reporting AUC values against topics using Logistic Regression, (b) Random Forest, and (c) Support Vector Machine classifier with 1:1 spread using interest+graph based features with 10% links known generated for 1,000 user data set.*

Table 5.21: *Comparing highest AUC values reported by Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM) classifiers with 1:1 spread using interest+graph based features with 10% links known generated using the ontology approach and LDA topic modeling approach over 1,000 user data set.*

Classifiers	Ontology Approach	LDA Topic Modeling Approach
LR	**0.6672(4)**	0.625(140)
RF	**0.6346(4)**	0.6292(90)
SVM	**0.7376(12)**	0.648(80)

presented here determine the performance of LR, RF and SVM classifiers.

Table 5.22: *Comparing highest AUC values reported by Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM) classifiers with 1:1 spread using interest+graph based features with 25% links known generated using the ontology approach and LDA topic modeling approach over 1,000 user data set.*

Classifiers	Ontology Approach	LDA Topic Modeling Approach
LR	**0.78(8)**	0.723(20)
RF	**0.8258(12)**	0.8024(60)
SVM	**0.7926(11)**	0.7684(80)

Similarly, under Table 5.22 and Figure 5.26 we compare the AUC values reported by the Logistic Regression, Random Forest and Support Vector Machine classifiers with 1:1 spread using interest+graph based features with 25% links known generated using ontology approach and interest+graph based features with 25% links known generated using LDA topic modeling technique. The table suggests that AUC values obtained using ontology approach outperformed AUC values obtained using LDA approach. Although, Logistic Regression shows an inconsistent performance it gives a higher AUC value for ontology approach as compare to LDA approach. Graphs generated plot the performance of LR, RF and SVM classifiers.

Under Table 5.23 and Figure 5.27 we compare AUC values reported by the Logistic Regression, Random Forest and Support Vector Machine classifiers using interest+graph based features with 50% links known generated using ontology approach and interest+graph based features with 50% links known generated using LDA topic modeling technique. Studying

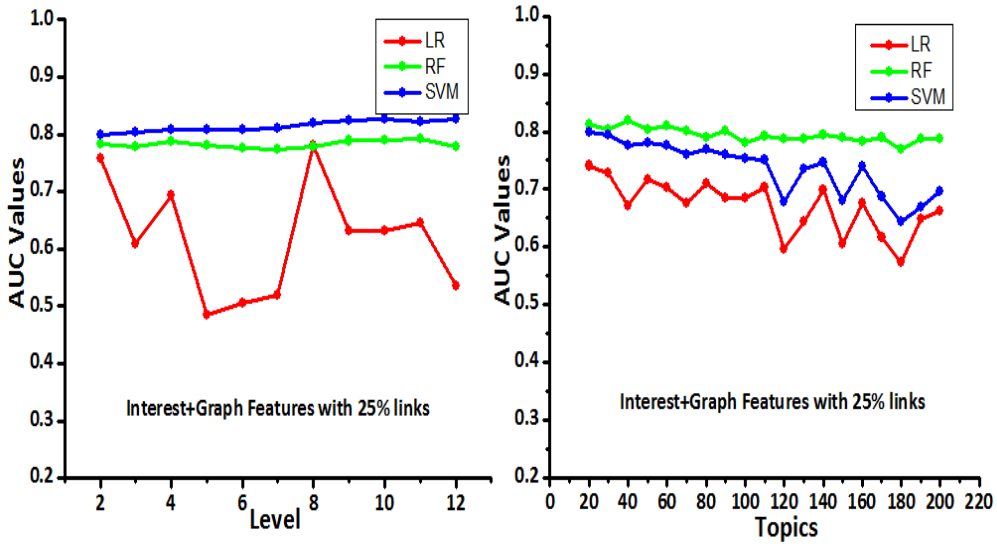

Figure 5.26: *Graph obtained for ontology approach by reporting AUC values against the number of levels and for LDA approach by reporting AUC values against topics using Logistic Regression, (b) Random Forest, and (c) Support Vector Machine classifier with 1:1 spread using interest+graph based features with 25% links known generated for 1,000 user data set.*

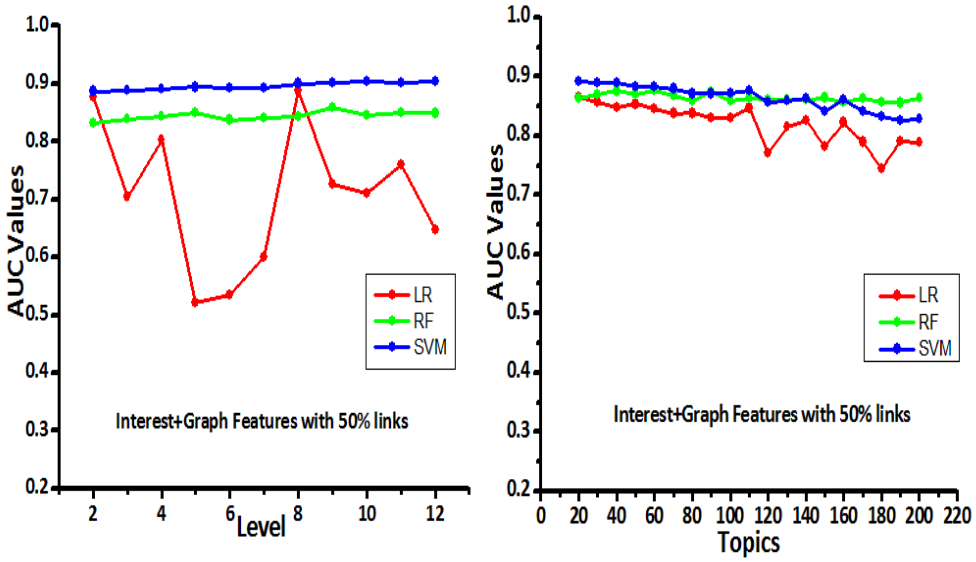

Figure 5.27: *Graph obtained for ontology approach by reporting AUC values against the number of levels and for LDA approach by reporting AUC values against topics using Logistic Regression, (b) Random Forest, and (b) Support Vector Machine classifier with 1:1 spread using interest+graph based features with 50% links known generated for 1,000 user data set.*

Table 5.23: *Comparing highest AUC values reported by Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM) classifiers with 1:1 spread using interest+graph based features with 50% links known generated using the ontology approach and LDA topic modeling approach over 1,000 user data set.*

Classifiers	Ontology Approach	LDA Topic Modeling Approach
LR	**0.8878(8)**	0.8594(20)
RF	**0.904(12)**	0.872(30)
SVM	0.8498(11)	**0.8808(60)**

this table we see that Logistic Regression and Random Forest classifiers when used with ontology approach outperformed LDA topic modeling technique, while, performance from Support Vector Machine for LDA approach outperformed the ontology approach.

Chapter 6

Conclusion & Future Work

The results obtained in Chapter 5 help us in answering a number of research related questions raised in Chapter 4. We discuss these questions and some of the limitations of ontology approach in Section 6.1. We discuss future work that can be done in order to improve and extend the ontology approach in Section 6.2.

6.1 Conclusion

In Chapter 4, we raised several research questions. The following are insights we gained based on the results of the experiments that we performed.

- Does the combination of *Graph-based Features* and *Interest-based Features* give better results than when *Graph-based Features* or *Interest-based Features* are used alone, when working with data sets containing more than 1000 users?

  We answer this question by making observations from Table 5.1, 5.2, 5.3, 5.4, 5.5, 5.6 and 5.7. Results obtained from 1000 user dataset suggested that for classifiers Logistic Regression, Support Vector Machine and Random Forest interest+graph features showed better performance as compared to interest features alone and graph features alone. However, Decision Tree (J48) classifier performed poorly when used with features generated from 1000 user dataset. For 2000 user datatset Support Vector Machine, Random Forest and J48 classifier showed an improvement in the performance

when interest+graph based features where used as opposed to interest based features alone and graph based features alone. Furthermore, for 3000 user dataset, in general all four classifiers performed better when used with interest+graph based features. A similar observation was shown by 4000 user dataset. Apart from that, J48 classifier showed an improvement in performance when the number of users in the dataset increases. Results obtained in Chapter 5 confirms our hypothesis, that in general, the performance of classifiers improves when presented with interest+graph based features as opposed to interest features alone and graph features alone immaterial of the size of the datasets.

- How do the classifiers perform over the data sets used in this work? Out of the four classifiers, which one performs the best?

  Observing Table 5.8, 5.9, 5.10 and 5.11, we see that except for Support Vector Machine classifier the performance of Logistic Regression, Random Forest and J48 improves with increase in the dataset size when used with graph based features with 10%, 25% and 50% links known. Similarly, observing results reported using interest based features generated at all ontology levels under Tables 5.12, 5.13, 5.14 and 5.15, we see that the performance of the Logistic Regression, Support Vector Machine, Random Forest and J48 classifiers improves with increase in the size of the datasets. Furthermore, observing results obtained for interest+graph based features with 10%, 25% and 50% links known from Tables 5.16, 5.17, 5.18 and 5.19, we see that the performance of the classifier improves with increase in the size of dataset.

  Comparing the overall performance of the classifiers for each data set, we see that Support Vector Machine gave best performance for 1000 and 4000 user datasets, followed by Logistic Regression, Random Forest and J48 classifiers. While, for 2000 user dataset best performance was shown by Logistic Regression classifier followed by Random Forest, Support Vector Machine and J48 classifiers. Furthermore, for 3000 user dataset best performance was shown by Logistic Regression classifier followed by

Support Vector Machine, Random Forest and J48 classifiers.

- How does the ontology approach considered in this thesis perform when the number of users is increased?

  From the above discussion, we suggest that the effectiveness of the ontology approach improves as the number of users increases in the dataset. However, the ontology approach becomes quickly intractable as we increase the number of users. This restricts us to evaluate the performance of ontology approach over datasets consisting of more than 4000 users of *LiveJournal* social network.

- Does the LDA based approach mentioned in [Parimi, 2010] perform better than the ontology approach considered in this thesis at the task of predicting friendship links in *LiveJournal* social network?

  Comparison between the results obtained by the two approaches have been reported in Chapter 5 under Section 5.3. We compare results obtained from Logistic Regression, Support Vector Machine and Random Forest classifiers with interest based features in Table 5.20, interest+graph features with 10% links known in Table 5.21, interest+graph features with 25% links known in Table 5.22 and interest+graph features with 50% links known in Table 5.23. Observing these tables, we see that LDA approach outperformed ontology based approach when only interest based features are considered. While, ontology based approach outperformed LDA approach using interest+graph features with 10% and 25% links for Logistic Regression, Support Vector Machine and Random Forest classifiers. Furthermore, ontology based approach outperformed LDA approach using interest+graph features with 50% links known for Logistic Regression and Random Forest classifiers, while LDA approach gave better performance over ontology approach for Support Vector Machine classifier using interest+graph based features with 50% links known.

Although, we showed that the performance of the classifiers improves at the task of predicting friendship links in *LiveJournal* social network, the ontology approach used in this work suffered from a major limitation. One of the biggest limitation of ontology approach was that it took a lot of time to generate interest based features as the number of users increases in the datasets. Furthermore, the ontology approach becomes quickly intractable as the number of users increases in the dataset.

6.2 Future Work

As a part of future work, we would like to improve the efficiency of the algorithm in order to compute interest based features and interest+graph features more efficiently, since the ontology approach [Haridas, 2009] becomes quickly intractable as the number of users increases in the data sets. Currently, in this work we have covered up to 4000 users of *LiveJournal* social network. However, in reality LiveJournal data consists of approximately 38,000 users. We would like to extend the ontology approach to cover complete *LiveJournal* data set.

Bibliography

W. Aljandal, W. Hsu, and V. Bahirwani. Validation-based normalization and selection of interestingness measures for association rules. 2008.

V. Bahirwani. Ontology engineering and feature construction for predicting friendship links and users interests in the live journal social network. Master's thesis, Kansas State University, 2008.

S. Burger. Ontology-based classification of unstructured information. 2010.

M. Ceci and D. Malerba. Classifying web documents in a hierarchy of categories: a comprehensive study, 2007.

B. Fitzpatrick. Live journal: online journal service. 1999.

L. Getoor. Link mining: a new data mining challenge. *SIGKDD Explor. Newsl.*, 5(1):84–89, 2003. ISSN 1931-0145. doi: http://doi.acm.org/10.1145/959242.959253.

M. Grobelnik and D. Mladenič. Simple classification into large topic ontology of web documents. 4:279–285, 2005.

M. Grobelnik, J. Brank, D. Mladenič, B. Novak, and B. Fortuna. Using dmoz for constructing ontology from data stream. In *Proceedings of the 28th International Conference on Information Technology Interfaces, Cavtat, Croatia*, 2006.

M. Haridas. Exploring knowledge bases for engineering a user interests hierarchy for social network applications. Master's thesis, Kansas State University, 2009.

W. H. Hsu, J. Lancaster, Paradesi M., and T. Weninger. Structural link analysis from

user profiles and friends networks: A feature construction approach. In *Proceedings of International Conference on Weblogs and Social Media*, pages 75–80, 2007.

L. Lu and T. Zhou. Link prediction in weighted networks: The role of weak ties. 2010.

T. M. Mitchell. *Machine learning*. McGraw-Hill Companies Inc., 1997.

D. Nowell and J. Kleinbergz. The link prediction problem for social networks. 2004.

R. Parimi. Lda based approach for predicting friendship links in live journal social network. Master's thesis, Kansas State University, 2010.

A. Patil. Homophily based link prediction in social networks. 2009.

H. Song, T. Cho, V. Dave, Y. Zhang, and L. Qiu. Scalable proximity estimation and link prediction in online social networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, 2009.

L. Tang and H. Liu. Graph mining applications to social network analysis. pages 487–531, 2010.

B. Taskar, M. Fai Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *Proceedings of the 17th Neural Information Processing Systems (NIPS)*, 2003.

G. Weiss and F. Provost. The effect of class distribution on classifier learning: An empirical study. 2001.

I. H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. J. Cunningham. Weka: practical machine learning tools and techniques with java implementations. In *Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems*, pages 192–196, 1999.