An Efficient Modified Common Neighbor Approach for Link Prediction in Social Networks

[1]Abhisek Chaturvedi, [2]Tapodhir Acharjee
[1]*Dept of Information Technology, Assam University, Silchar-788011, India*
[2]*Dept of Information Technology, Assam University, Silchar-788011, India*

Abstract: *Link prediction in social networks aims at estimating the likelihood of the appearance of a new link between two nodes, based on the existing links and the attributes of the nodes. Many methods for link prediction problem in social networks have been proposed in literature. We especially analyze the shortcomings of common neighbor leading method. Accordingly we generate a new modified common neighbor approach for link prediction in social networks. Our approach efficiently works under the integrated analysis of features along with topological structure of a social network. As a co-authorship network is a true social network, we have considered the co-authorship networks for verifying the effectiveness of the existing leading methods as well as our proposed link prediction method. We have implemented the leading methods as well as our proposed method on two different data domains of co-authorship networks obtained from author lists of papers at five sections of Physics e-Print arXiv, www.arXiv.org. In the first data domain, the papers in the periods (1994 – 1996) and (1997 – 1999) are taken as the training set and testing set respectively. Similarly the papers in the periods (2007 – 2009) and (2010 – 2012) are taken as the training set and testing set for the second data domain. Experimental results show that all the methods are found to perform much better over the random predictor. Again we find that our modified common neighbor approach outperforms over all the existing leading methods considered.*

Keywords: *co-authorship networks, common neighbor, modified common neighbor approach, link prediction, random predictor*

Submitted date 14 June 2013 **Accepted Date: 19 June 2013**

I. Introduction

Liben-Nowell and Kleinberg [4,5] explain link prediction problem as: *"Given a snapshot of a social network at time t, we seek to accurately predict the edges that will be added to the network during the interval from time t to a given future time t'."*

In our daily lives we often see many complex networks around us (e.g. Internet, WWW, Transportation Network, Social Network etc.). The social network may be viewed as a graph where all the nodes or actors are individuals or organizations and links are contacts (interactions, relationships or collaborations) which show the proximity or similarity among those individuals or organizations. Natural examples of social networks include the set of all scientists in a particular discipline, with edges joining pairs who have co-authored papers; the set of all employees in a large company, with edges joining pairs working on a common project; a collection of business leaders, with edges joining pairs who have served together on a corporate board of directors.

In past social networks have been studied by many researchers for different purposes. The challenging problem to deal with these networks has been their highly dynamic nature; they grow and change quickly over time through the addition of new edges which signifies the appearance of new interactions in the underlying social structure. Understanding the mechanism by which they evolve is a fundamental question that is still not well understood by us, and it forms the motivation for our work here on link prediction problem. Recently the link prediction in complex networks has attracted increasing attention from computer scientists and physicists. Link prediction aims at estimating the likelihood of the appearance of a new link between two nodes, based on the existing links and the attributes of the nodes. For example, classical information retrieval can be viewed as predicting missing links between words and documents, and the process of recommending items to users can be considered as a link prediction problem in the user-item bipartite networks. Link prediction problem can be categorized into two categories: one is the prediction of existing yet unknown links (e.g. criminal/terrorists networks) and the other is the prediction of links that may appear in the future of evolving networks (e.g., co-authorship networks). Here, we focus on second category of link prediction problem.

Many methods for link prediction problem in social networks have been proposed in literature. Here we focus on existing leading methods and analyze the shortcomings of some of these methods. Accordingly we have come out a few new methods for link prediction in social networks. Our methods are based on the integrated analysis of features along with topological structure of social networks. As a co-authorship network is a true social network, we have considered the co-authorship networks for verifying the effectiveness of the existing leading as well as our proposed methods. We have implemented the all the methods on two different data domains of co-authorship networks obtained from author lists of papers at five sections of Physics e-Print arXiv, www.arXiv.org. In the first data domain, the papers in the periods (1994 – 1996) and (1997 – 1999) are taken as the training set and testing set respectively. Similarly the papers in the periods (2007 – 2009) and (2010 – 2012) are taken as the training set and testing set for the second data domain. Here for modeling of all of these co-authorship networks, we have proposed an efficient algorithm. Experimental results show that all the methods are found to perform much better over the random predictor on both the data domains. Again we find that our proposed hybrid approach that is based on the common neighbor-wise combination of our two proposals (modified Common neighbor and modified Adamic/Adar) gives better results than the existing leading methods considered on both the data domains.

Rest of paper is organized as follows. In Section 2, we will focus on the literature survey related to link prediction problem. Section 3 finds a short description of existing leading link prediction methods. Our proposed link prediction methods are furnished in section 4. Section 5 is concerned with the implementation and results. The paper is concluded in section 6.

II. Literature Survey

With the best of our knowledge, the first work on link prediction problem in social networks was done by Liben-Nowell and Kleinberg [4]. They tested the predictive power of some proximity metrics, including Common neighbours, Adamic/Adar, Katz measure. They further worked on this problem in [5] with more descriptive analysis of same proximity metrics as in [4]. Here in both the works [4][5], their hypothesis was that link prediction could be performed from topological analysis alone. Some authors in [6] studied various aspects like modeling, clustering and ranking in co-authorship networks induced by IEEE, ACM and joint IEEE/ACM digital library conferences. Some other authors in [7] explored learning based approach where the predictors are learned using a combination of features and then the performance is evaluated. In 2004, Popescul and Ungar enhanced link prediction of author/document bipartite networks by using clustering [14]. In [15], Zhou and Scholkopf approached three related graph problems (classification, ranking and link prediction) in a new way. In [11], Murata and Moriyasu proposed three weighted similarity indices, as variants of the Common Neighbors, Adamic-Adar and Preferential Attachment indices, respectively. Some authors in [8] measured the performances of weighted and unweighted versions of Common neighbor, Adamic-Adar and Resource Allocation on real social, technological and biological networks, and found that sometimes the weighted indices perform even worse than unweighted indices. Zhou et al. [9, 10] studied nine well-known local similarity indices on six real networks extracted from disparate fields, as well as proposed two new local indices. More related is the work on link prediction using multiple structural attributes and research titles by Pavlov- Ichise and Wohlfarth- Ichise [12, 13]. They introduced structural features of the graph trying to capture the network structure effectively. Then, they used supervised learning techniques for link prediction. In [17], Huang, Li and Chen investigated the use of link prediction to improve collaborative filtering in recommender systems. In [16], link prediction was used to design a system that recommended new academic links for researchers at a computer science conference and received feedback through a survey. In [18], Zhu used link prediction to determine what web page a user was next likely to visit in order to improve the navigation and efficiency of a site. A few other link prediction papers are summarized by Getoor and Diehl in [19].

III. Existing Leading Link Prediction Methods & Analysis

Many methods have been proposed in literature for link prediction in social networks. Some leading methods [4][5][11] are : Common neighbor, Jaccard coefficient, Adamic/Adar, Katz Measure, Rooted Page Rank, Low Rank Approximation, Unseen bigram, Clustering. Here we analyze the shortcomings of methods by making special intuition for common neighbor method as follows:

- The major problem not only with common neighbor approach but with all of other methods is that they just focus on topological structure alone to predict the links in social networks. There is no node-similarity based approach is followed so need to be modified.
- Traditional weighted Common neighbor method [11] which is not considered to be better than unweighted Common neighbor [4-5], just takes simple addition of all of the weights of common nodes' links with nodes *X* and *Y*, no clear logic is shown for this addition so need to be modified.
- Bigram approach [4-5] is not truly defined for link prediction in social networks. The actual bigram approach refers to find semantic relations between any two nodes *X* and *Y* (e.g. authors) based on their features' selection. This limitation for link prediction highly motivates us to work on features for prediction in social networks.\

• We can find that till now, no one has provided an unsupervised efficient approach for link prediction in social networks under integrated analysis of features along with topological structure of social networks.

IV. Proposed Link Prediction Method

After analyzing the shortcomings of common neighbor and bigram existing leading methods in section 3, here we provide a modified common neighbor method which works under integrated analysis of features along with topological structure of social network. Our approach can be well understood under our suggested normalized technique that is shown below:

$$w(X,Y) = 2k/(n_1+n_2) \quad \ldots\ldots\ldots\ldots\ldots(1)$$

Here n_1 and n_2 denote total number of features of nodes *X* and *Y* respectively; and *k* denotes the number of common features between any two connected nodes *X* and *Y*.
It generates the weights for each link in present structure which lie between 0 and 1.

4.1 Modified Common Neighbor Method

Common neighbors between nodes *X* and *Y* play big roles in prediction of link between them. These common neighbors can be called 'brokers' or in concept of social network 'structural holes' which have the information about all those nodes which are connected with these nodes ('brokers'). The level of information that these nodes (brokers) have about other nodes connected with them, depends on the strength of their relationships. So our proposed method concentrates on the strength of links (relationships) of each common node ('broker') with nodes *X* and *Y*.
The formula is given as follows:

$$Score_{MCN(f)}\ (X, Y) = \sum_{n=0}^{n=N} w_1 \times w_2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Here, *N*- total number of common neighbors and w_1, w_2 indicate strength of connections of each respective common node with *X* and *Y*.
If we talk more specifically then our method generates the actual contribution score with respect to strength of links of each common node with *X* and *Y*. Here we can write for each common node: 0<= contribution score <=1.
We can understand with the help of some examples which highly motivate us. Let us take *X* and *Y* any two unconnected nodes and here Z is a common node between them.
1. If w(*X*,*Z*)=1 and w(*Z*,*Y*)=0.24 then generated score will be 0.24; Here 1 indicates that both *X* and *Z* are fully similar in behavior so there would be similar likes and dislikes. And now if we focus on the strength of link between *Y* and *Z* then we find, it is not too good that means at a great extent both of them are of different behavior. In other words, we can say that relationship between *Z* and *Y* is not good. Now if *Z* does not like *Y* then *X* will also not like *Y* at same level because *X* and *Z* are fully similar in nature. So the weight of predictable link will remain 0.24. In this way, we can say that the generated score is a true score that lies between 0 and 1 for each respective common node between *X* and *Y*.
2. If w(*X*, *Z*)=1 and w(*Z*,*Y*)=1 then generated score will be 1. Obviously, persons in pair (*X*, *Z*) and in pair (*Z*, *Y*) are of similar behaviors respectively and know well about one another because of full matching of all features. So *X* and *Y* will also be of similar behavior and in this way, strength of predictable link will be very good (say, 1).
3. If w(*X*, *Z*)=0.76 and w(*Z*,*Y*)=0.0 then generated score will be 0. Obviously, we can say that *Z* and *Y* will not be connected in present structure and can be unknown to one another at all so it cannot be a good idea to make prediction between *X* and *Y* in this case under integrated analysis of features along with topological structure of networks.
This proposal holds three key intuitions for taking the summation of each product of weights of each common node's links with nodes *X* and *Y*:
➢ If both the links are strong then predicted link will be strong or almost strong (depending on strength of links how much they are strong). (e.g. if w_1=1 and w_2=0.95 then *score(X,Y)*=0.95).
➢ If both the links are weak then predicted link will be weak or very weak (depending on the strength of links how much they are weak). (e.g. if w_1=0.3 and w_2=0.2 then *score(X,Y)*=0.06).
➢ If one link is strong and other is weak then predicted link will also be weak or very weak (depending on the strength of links how much a strong link is strong and a weak link is weak). (e.g. if w_1=1 and w_2=0.3 then *score(X,Y)*=0.3).

V. Implementation and Results

As a co-authorship network is a true social network and many readers are likely to be interested on such a network for personal as well as scientific reasons, so we consider co-authorship networks for verifying the effectiveness of the existing as well as our new method.

5.1 Formulation of Link Prediction Problem

Suppose we have a social network $G = < V, E >$ in which each edge e $= < u, v >$ represents an interaction between u and v that took place at a particular time *t* (e). We record multiple interactions between *u* and *v* as parallel edges, with potentially different time-stamps. For two times $t < t'$, let $G[t, t']$ denote the subgraph of *G* consisting of all edges with a time-stamp between t and t'. Here, then, is a concrete formulation of the link prediction problem. We choose four times $t_0 < t_0' < t_1 < t_1'$, and give an algorithm access to the network $G[t_0, t_0']$; it must then output a list of edges, not present in $G[t_0, t_0']$, that are predicted to appear in the network $G[t_1, t_1']$. We refer to $[t_0, t_0']$ as the *training interval* and $[t_1, t_1']$ as the *test interval.*

Of course, social networks grow through the addition of nodes as well as edges, and it is not sensible to seek predictions for edges whose end points are not present in the training interval. Thus in evaluating link prediction methods, we will generally use two parameters $k_{training}$ and k_{test}, and define the set Core to be all nodes incident to at least $k_{training}$ edges in $G[t_0, t_0']$ and at least k_{test} edges in $G[t_1, t_1']$. We will then evaluate how accurately the new edges between elements of Core can be predicted.

5.2 Preprocessing Work on Datasets

5.2.1 Data Source

Five sections of the physics e-Print arXiv at Cornell University's digital library website: www.arxiv.org. *ii) Information about Required Datasets:* We work on five co-authorship networks: astro-ph (astrophysics), cond-mat (condensed matter), gr-qc (general relativity and quantum cosmology), hep-ph (high energy physics-phenomenology), and hep-th (high energy physics- theory), on two different data domains where in the first data domain, the papers in the periods (1994 – 1996) and (1997 – 1999) are taken as the training set and testing set respectively. Similarly the papers in the periods (2007 – 2009) and (2010 – 2012) are taken as the training set and testing set for the second data domain.

5.2.2 Capturing of Data

We simply access the Cornell University's digital library website: www.arxiv.org on a web browser and we find different sections of the physics e-Print arXiv at this site. Then we find out our required five co-authorship networks and take the raw data for these networks on both the data domains. This raw data of each network on both the data domains holds papers' information (say, paper_id, title, authors and journal if present). Finally for both data domains, we place raw data of each network for training and test periods in different text tiles.

5.2.3 Structuring of Data

As the whole raw data is in unstructured format so our next task is to provide a proper structure of these raw datasets of each network on both the data domains. We convert the raw data in a proper structural format (here in each text file having data in structured format, each row holds: paper ids, papers' titles, authors and journal if available otherwise null)) for our whole work on two different data domains.

5.2.4 Core Information

In our experiments on the arXiv, we can identify which authors are active throughout the entire period on the basis of the number of papers published. Thus here we define the set Core to consist of all authors who have written at least $k_{training}$ = 3 papers during the training period and at least k_{test} = 3 papers during the test period for each network at first of first dada domain (1994-1999) [Table: 1] and then of second domain (2007-2012) [Table: 2]. Here we choose '3' that can give more possibility to stay the authors in testing period also.

Table 1: Core data of five co-authorship networks at first data domain.

Networks	Training period			Core data		
	authors	papers	edges	authors	$\|E_{old}\|$	$\|E_{new}\|$
astro-ph	5308	5806	20913	1564	6162	9086
cond-mat	5783	6628	11687	1315	2186	2362
gr-qc	2139	3286	2900	457	488	504
hep-ph	5433	10257	21694	1771	6142	5564
hep-th	5253	9728	7992	1423	2293	2550

Table 2: Core data of five co-authorship networks at second data domain.

Networks	Training period			Core data		
	authors	papers	edges	authors	$\|E_{old}\|$	$\|E_{new}\|$
astro-ph	28815	33432	504315	11040	280129	368635
cond-mat	32964	30907	166046	7346	51336	60453
gr-qc	6932	9380	34081	1650	6700	6602
hep-ph	10751	16247	90802	3343	23476	26936
hep-th	8909	15213	20502	2798	7011	6913

5.3 Implementation

Now we describe our implementation setup more specifically. Let us take any one of five co-authorship networks in the first data domain (see Table 1), and denote the subgraph $G[1994,1996]$ on the training interval by G_{collab} := <A,E_{old}> and use E_{new} to denote the set of edges <u,v> such that <u, v>ϵA(indicating core authors), and u,v co-author a paper during the test interval but not the training interval - these are the new interactions we are seeking to predict. In a similar way, we can do for the subgraph $G[2007,2009]$ of any one of these five networks in second data domain(see Table 2).

Now at this subgraph G_{collab} := <A,E_{old}> for each co-authorship network in both the data domains(see Table 1 & 2), at first we implement all the leading methods described in section 3 and then for implementing our proposed link prediction method described in section 4, we focus on titles of papers for getting information about features of authors/scientists (say, attributes). For this work we use Google stop words vocabulary[1] and Porter stemmer algorithm [24]; and generate the informative keywords (stems) for each author. After finding these keywords for each author, we implement our proposed method on G_{collab} under integrated analysis of features along with topological structure of social networks.

Further we evaluate each link prediction methods as follows: Each method p that we consider outputs a ranked list L_p of pairs in $\mathrm{A} \times A - E_{old}$; these are predicted new collaborations, in decreasing order of confidence. For our evaluation, we focus on the set Core (see Table: 1 & 2 in section 5), so we define $E^*_{new} = E_{new} \cap (Core \times Core)$ and n =$|E^*_{new}|$. (See Table: 3 & 4) Our performance measure for predictor p is then determined as follows: from the ranked listL_p, we take the first n pairs in $Core \times Core$, and determine the size of the intersection of this set of pairs with the set E^*_{new} .

For whole implementation work, we have used the machine with configuration 48 GB RAM, Intel Xeon Processor 12 Core 2.4 GHz Speed and our programming platform has been NetBeans IDE 7.1.2 on Windows 7 OS.

[1] Online SEO Guide:Google Stop Words Vocabulary: http://www.link-assistant.com/seo-stop-words.html

Table 3: Finding E^{*}_{new} for evaluating each link predicting method on each of five co-authorship networks in first data domain

Existing LP Methods	**astro-ph**	**cond-mat**	**gr-qc**	**hep-ph**	**hep-th**	**Total**
E_{new}	*9086*	*2362*	*504*	*5564*	*2550*	*20066*
RP	46	5.6	1	13.4	4.6	70.6
CN(unwgtd)	***477***	***74***	***11***	***189***	***91***	***842***
CN(wgtd)	417	81	8	200	101	807
AA(unwgted)	***526***	***87***	***16***	***247***	***93***	***969***
AA(wgted)	247	89	9	182	96	623
JC	435	68	6	249	81	839
KM_1(wgtd)	382	92	7	176	97	754
KM_2(wgtd)	382	91	7	176	97	753
KM_3(wgtd)	375	89	7	174	95	740
KM_1(unwgtd)	447	77	11	186	99	820
KM_2(unwgtd)	447	77	11	186	99	820
KM_3(unwgtd)	409	74	11	179	99	772
SimRank (γ=0.8)	405	40	11	167	72	695
RPR(α=0.15)	435	70	6	209	86	806
RPR(α=0.50)	410	83	9	196	102	700
LRA(1024)+CN	462	70	12	162	75	841
LRA(512)+CN	399	60	12	151	64	686
LRA(256)+CN	297	54	13	101	62	527
KC(β:.001,ρ:.10)	382	92	7	175	95	751
KC(β:.001,ρ:.15)	385	92	7	177	96	757
KC(β:.001,ρ:.20)	386	92	7	183	97	765
KC(β:.001,ρ:.25	379	91	7	184	100	661
UB(δ=4)+CN	414	59	8	214	82	777
UB(δ=5)+CN	437	60	4	215	74	790

Table 4: Finding E^{*}_{new} for evaluating each link predicting method on each of five co-authorship networks in second data domain

$\boldsymbol{E^{*}_{new} = E_{new} \cap (Core \times Core}$	**astro-ph**	**cond-mat**	**gr-qc**	**hep-ph**	**hep-th**
$\boldsymbol{(Core \times Core)}$	59819039	25793119	1119050	5275394	3384699
$\boldsymbol{E_{new}}$	368635	60453	6602	26936	6913
$\boldsymbol{E^{*}_{new}}$	246524	40192	3395	19426	3911

5.4 Implementation Results

Now we show the experimental results of existing and proposed link predicting methods at first on each of five co-authorship networks in first data domain (See Table:5 and Table:6) and then on each of five co-authorship networks in second data domain (See Table:7 and Table:8). Here for measuring the overall performances of existing and proposed methods, we sum the number of correct predictions made by respective method on each of five networks. For better and meaningful evaluation of all link prediction methods' quality, we use as our baseline a random predictor which simple randomly selects pairs of authors who did not collaborate in the training interval. Here we also avoid the biasness of random predictor's results for each respective network on both the data domains.

5.4.1 Results on First Data Domain (1994-1999)

Results for each of five co-authorship networks on first data domain are as follows:

5.4.1.1 Experimental Results of Existing LP Methods over Random Predictor

Experimental results of existing link prediction methods over random predictor are given in Table 5. Here we give abbreviations for some methods which are as follows: RP: Random Predictor, CN: Common neighbor, AA: Adamic/Adar, JC: Jaccard coefficient, KM(KM_1, KM_2, KM_3): Katz measure with β=0.0005, 0.005, 0.005 respectively, RPR: Rooted page rank, LRA: Low rank approximation, KC: Katz clustering, UB: Unseen bigram), wgtd: weighted, unwgtd: unweighted. Here we place E_{new} in Table 5 for better visualization of predictors' performances. Methods with bold italic entries give better results than other existing leading methods considered.

$\boldsymbol{E^*_{new} = E_{new} \cap (Core \times Core)}$	astro-ph	cond-mat	gr-qc	hep-ph	hep-th
$\boldsymbol{(Core \times Core)}$	11,45241	704080	69637	1392486	784592
$\boldsymbol{E_{new}}$	9086	2362	504	5564	2550
$\boldsymbol{E^*_{new}}$	5595	1207	186	3125	1373

5.4.1.2 Experimental Results of Proposed LP Methods over Random Predictor

Experimental results of proposed link prediction method over random predictor are given in Table 6. Here we give abbreviation for Modified Common Neighbor as MCN. Similar to Table 5, here also we place E_{new} for better visualization of our proposed link prediction method's performance.

Table 6: Performance of proposed link predicting method over random predictor

Proposed LP Methods	astro-ph	cond-mat	gr-qc	hep-ph	hep-th	Total
E_{new}	*9086*	*2362*	*504*	*5564*	*2550*	*20066*
RP	46	5.6	1	13.4	4.6	70.6
MCN	513	100	13	285	111	1022

5.4.2 Results on Second Data Domain (2007-2012)

Results for each of five co-authorship networks on second data domain are as follows:

5.4.2.1 Experimental Results of Existing LP Methods over Random Predictor

Experimental results of existing link prediction methods over random predictor are given in Table 7. Here we give abbreviations for some methods which are as follows: RP: Random Predictor, CN: Common neighbor, AA: Adamic/Adar, JC: Jaccard coefficient, KM (KM_1, KM_2, KM_3): Katz measure with β=0.0005, 0.005, 0.005 respectively, RPR: Rooted page rank, LRA: Low rank approximation, KC: Katz clustering, UB: Unseen bigram), wgtd: weighted, unwgtd: unweighted. Here we place E_{new} in Table for better visualization of predictors' performances. Methods with bold italic entries give better results than other existing methods.

Table 7: Performances of existing link predictors over random predictor

Existing LP Methods	astro-ph	cond-mat	gr-qc	hep-ph	hep-th	Total
E_{new}	*368635*	*60453*	*6602*	*26936*	*6913*	*469539*
RP	1519	94.2	17.2	94.4	8.4	1733.2
CN(unwgtd)	***25752***	***4114***	***352***	***1443***	***196***	***31857***
CN(wgted)	20093	2876	276	1337	221	24803
AA(unwgtd)	***26706***	***4295***	***393***	***1659***	***227***	***33280***
AA(wgted)	21104	2751	229	1277	220	25581
JC	25207	2252	284	1293	188	29224
KM(wgted)	20113	2890	280	1332	223	24838
KM(wgted)	20101	2888	280	1339	222	24830
KM(wgted)	20085	2881	276	1328	222	24792
KM(unwgted)	24167	3905	330	1205	191	29798
KM(unwgted)	23993	3776	310	1176	190	29355
KM(unwgted)	23908	3442	33	698	152	28233
SimRank γ=0.8	23124	3805	183	1287	192	28591
RPR(α=0.15)	22610	3901	153	1198	151	28013
RPR(α=0.50)	21008	4103	170	1177	143	26601
LRA(1024)+CN	25641	4131	315	1219	130	31436
LRA(512)+CN	24890	4043	271	992	101	30297
LRA(256)+CN	23078	4053	352	899	100	28482
KC(β:.001,ρ:.10	20113	2890	283	1332	222	24840
KC β:.001,ρ:.15	20317	2932	287	1341	223	25100

KC β:.001,ρ:.20	20319	2932	287	1345	226	25109
KC β:.001,ρ:.25	20321	2935	284	1349	223	25112
UB(δ=4)+CN	21146	2094	345	1399	189	25173
UB(δ=5)+CN	21713	2368	334	1354	183	25952

5.4.2.2 Experimental Results of Proposed LP Method over Random Predictor

Experimental results of proposed link prediction method over random predictor are given in Table 8. Here we give abbreviation for Modified Common Neighbor as MCN. Here also we have place E_{new} for better visualization of our proposed link prediction method's performance.

Table 8: Performance of proposed link predicting method over random predictor

Proposed LP Methods	astro-ph	cond-mat	gr-qc	hep-ph	hep-th	Total
E_{new}	*368635*	*60453*	*6602*	*26936*	*6913*	*469539*
RP	1519	94.2	17.2	94.4	8.4	1733.2
MCN	26728	4423	362	1802	251	33566

5.4.3 Results based on Execution Time

Here we show the results for execution time taken by our proposed method over existing common neighbor method. Here these results are shown over gr-qc network in first data domain (1994-1999). In similar way, we can show for other networks also. Here our intuition in finding the execution time has been to check whether our method is efficient with respect to execution time over existing CN method or not (See Table 9).

Table 9: Execution time analysis, Here, CN: Common neighbor, MCN: Modified Common Neighbor.

Serial no.	Existing Methods	Proposed Method	Execution Time(sec)
1.	CN(unwgted)	-	*2.746*
2.	-	MCN	10.591

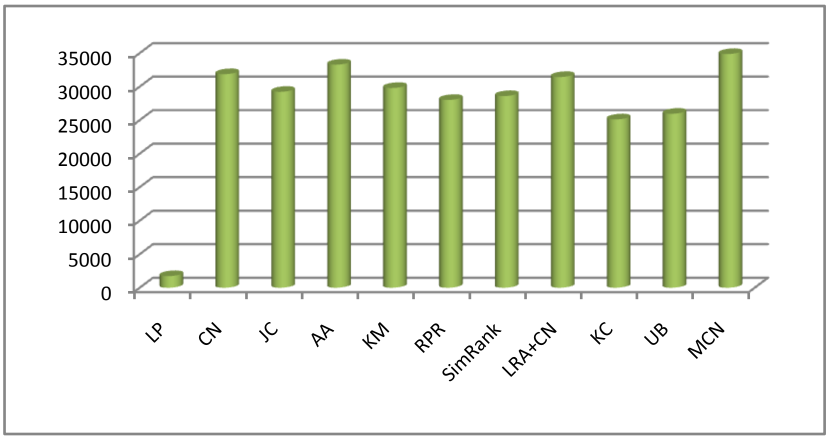

Figure 1: Correct Predictions vs Existing methods and proposed approach on First Data Domain (1994-99). Here, RP: Random Predictor, CN: Common Neighbor, JC: Jaccard Coefficient, AA: Adamic/Adar, KM: Katz Measure with β=0.0005, RPR: Rooted Page Rank with α=0.15, LRA: Low Rank Approximation with rank-1024, KC: Katz Clustering with ρ=0.20, UB: Unseen Bigram with δ=4, MCN: Modified Common Neighbor.

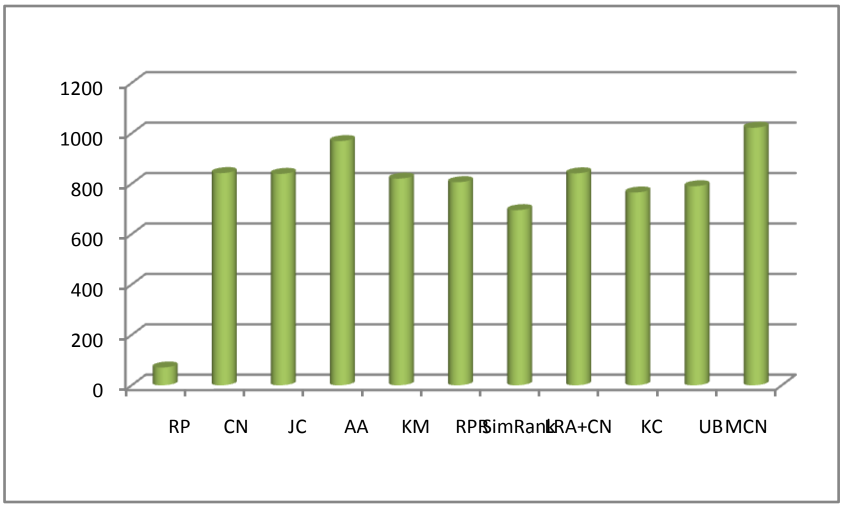

Figure 2: Correct Predictions vs Existing methods and proposed approach on Second Data Domain (2007-12). Here, RP: Random Predictor, CN: Common Neighbor, JC: Jaccard Coefficient, AA: Adamic/Adar, KM: Katz Measure with β=0.0005, RPR: Rooted Page Rank with α=0.15, LRA: Low Rank Approximation with rank-1024, KC: Katz Clustering with ρ=0.25, UB: Unseen Bigram with δ=5, MCN: Modified Common Neighbor.

VI. Conclusions

Experimental results show that all the existing leading methods considered and our proposed method are found to perform much better over the random predictor on both the data domains. Again we find that our approach that works under integrated analysis of features along with topological structure of social networks, gives better results than the existing leading methods considered on both the data domains.

Our suggestions for further work are as follows:

1. Here in our work, we focus on titles of papers to find the required features of authors. Anyone else can include other semantic information (e.g. abstracts, journals etc.) and can improve the performance of our proposed link prediction method.
2. Related to core information on each co-authorship network, we focus only on those authors who write at least three papers in training and testing periods respectively. We can do our task by taking authors with at least two papers or greater than three papers on large networks.
3. As our method has been time-consuming than existing leading methods considered (See Table 9 in section 5); so need to be modified for quick predictions on very large datasets.
4. Here we focus on co-authorship networks as these networks are true social networks. We can make an intuition for applying our method on other social networks like friendship networks, employees' networks in organizations.

References

[1] Salton G. and McGill M. J., Introduction to Model Information Retrieval (McGraw-Hill, Auckland) 1983.

[2] Zhou T., Ren J., Medo M. and Zhang Y.C., Bipartite network projection and personnel recommendation, Phys.Rev. E, 76 (2007) 046115.

[3] Zhou T., Jiang L.L., Su R.Q. and Zhang Y.C., Effect of initial configuration on network-based recommendation, EPL 81 (2008) 58004.

[4] Liben-Nowell, D., Kleinberg, J.: The link prediction problem for social networks. In: Proceedings of the twelfth international conference on Information and Knowledge Management, pp. 556-559.ACM Press, New York (2003).

[5] Liben-Nowell D. and Kleinberg J., J. Am. Soc. Inf.Sci. & Technol., 58 (2007) 1019.

[6] Xiaoming Liu, Johan Bollen, Michael Nelson, Herbert Van De Sompel. Co-authorship networks in digital library research community, Journal of Information Processing and Management, volume 41, issue 6, 2005.

[7] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, Mohammed Zaki. Link Prediction Using Supervised Learning, In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security, 2006.

[8] Linyuan Lšu and Tao Zhou, Link prediction in weighted networks: The role of weak ties, EPL, 89 (2010) 18001.

[9] Zhou T., Lšu L. and Zhang Y.-C., Eur. Phys. J. B, 71(2009) 623.

[10] Lšu L., Jin C.-H. and Zhou T., Phys. Rev. E, 80 (2009)046122.

[11] Murata T. and Moriyasu S., Link prediction of social networks based on weighted proximity measures, in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (ACM Press, New York) 2007.

[12] Pavlov, M., Ichise, R.: Finding experts by link prediction in co-authorship networks. In: Proceedings of the 2nd International Workshop on Finding Experts on the Web with Semantics (2007).

[13] Wohlfarth T., Ichise, R.: Semantic and Event-Based Approach for Link Prediction, in Proceedings of the7th International Conference on Practical aspects of Knowledge Management 50-61 (2008).

[14] Popescul, A and Ungar, LH 2004, 'Cluster-based Concept Invention for Statistical Relational Learning', Proceedings of Conference Knowledge Discovery and Data Mining (KDD-2004), 22-25 August 2004, viewed 12 June 2006.

[15] Zhou, D & Scholkopf, B 2004, 'A regularization framework for learning from graph data', Proceedings of Workshop on Statistical Relational Learning at International Conference on Machine Learning, Banff, viewed 12 June 2006.

[16] Farrell, S, Campbell, C & Myagmar, S 2005, 'Relescope: An Experiment in Accelerating Relationships', paper presented at Conference on Human Factors in Computing Systems, 2-7 April 2005, Portland.

[17] Huang Z., Li X. and Chen H., Link prediction approach to collaborative filtering, in Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries(ACM Press, New York) 2005.

[18] Zhu, J 2003, 'Mining Web Site Link Structures for Adaptive Web Site Navigation and Search', PhD thesis, University of Ulster, viewed 12 June 2006.

[19] Getoor and C. P. Diehl, “Link mining: a survey,” SIGKDD Explor. Newsletter, vol. 7, no. 2, pp. 3–12, 2005.

[20] Leo Katz. A new status index derived from sociometric analysis. Psychometrika, 18(1):39-43, March 1953.

[21] Han Hee Song, Tae Won Cho, Vacha Dave, Yin Zhang, Lili Qiu: Scalable proximity estimation and link prediction in online social networks. Internet Measurement Conference 2009: 322-335.

[22] Sergey Brin and Lawrence Page, The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 30(1-7):107-117, 1998.

[23] Glen Jeh and Jennifer Widom, SimRank: A measure of structural-context similarity, In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 271-279, July 2002.

[24] M.F. Porter, An algorithm for suffix stripping, Program, 14(3) pp 130−137, 1980.