Neurocomputing 172 (2016) 71–83

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Influence of edge weight on node proximity based link prediction methods: An empirical analysis

Niladri Sett*, Sanasam Ranbir Singh, Sukumar Nandi

Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India

ARTICLE INFO

Article history:
Received 1 November 2013
Received in revised form
22 July 2014
Accepted 5 November 2014
Available online 23 May 2015

Keywords:
Link prediction
Link weight
Node proximity
Strength of tie

ABSTRACT

Tie weight plays an important role in maintaining cohesiveness of social networks. However, influence of the tie weight on link prediction has not been clearly understood. In few of the previous studies, conflicting observations have been reported. In this paper, we revisit the study of the effect of tie weight on link prediction. Previous studies have focused on additive weighting model. However, the additive model is not suitable for all node proximity based prediction methods. For understanding the effect of weighting models on different prediction methods, we propose two new weighting models namely, *min-flow* and *multiplicative*. The effect of all three weighting models on node proximity based prediction methods over ten datasets of different characteristics is thoroughly investigated. From several experiments, we observe that the response of different weighting models varies, subject to prediction methods and datasets. Empirically, we further show that with the right choice of a weighting model, weighted versions may perform better than their unweighted counterparts.

We further extend the study to show that proper tuning of the weighting function also influences the prediction performance. We also present an analysis based on the properties of the underlying graph to justify our result. Finally, we perform an analysis of the *weak tie theory*, and observe that unweighted models are suitable for inter-community link prediction, and weighted models are suitable for intra-community link prediction.

1. Introduction

Given a social network graph, the task of link prediction problem can be broadly categorized into two (i) prediction of existing links, but yet unknown [1] and (ii) prediction of non-existing links, but likely to appear in the future [2]. Though, originally it started with social networks [2] and biological networks [3], recently the problem has attracted many other domains such as information retrieval [4], where the problem is to predict the missing relationship between words and documents from a word-document graph; and recommendation system [5], where the problem is to predict relationship between products and users. Initial studies [2,6,7] in link prediction methods such as *Common Neighbor* (*CN*), *Jaccard's coefficient* (*JC*), *Adamic/Adar* (*AA*), *Resource Allocation* (*RA*) have explored the topological characteristics of graphs by performing local analysis on node proximity. Majority of studies on the local analysis based link prediction methods consider unweighted graphs. Links in typical social networks, such as message passing, co-authorship and friendship are weighted in nature i.e., links are characterized by *strong* and *weak* ties [8]. It is natural to take tie weights into consideration for the link prediction problem. Traditionally, tie weights are represented by the frequency of interaction between two actors. A different way of representing the characteristics of the ties is also available in the literature [9], where the authors have used the content of a tie to improve the community detection performance. However, influence of tie weights on the local analysis based link prediction methods is not clearly understood. Few studies have been reported in the literature. First of such study has been reported in [10], where the authors have observed positive influence of weight on the above prediction methods. However later in studies [11,12], conflicting results are observed, where the former observes a negative influence and the later observes a positive influence. In all these studies, an additive weighting model has been used. It has not been thoroughly analyzed whether the reported weighting model in its current form is the best estimate of incorporating weights. This has motivated us to propose different forms of weighting models and investigate their performance on prediction methods and datasets. Like [10–12], this paper focuses only on the local analysis based link prediction methods.

From several experiments using ten datasets constructed from eight social networks, it is observed that different weighting models respond differently on prediction methods and datasets.

* Corresponding author. Tel.: +91 9678554481.
E-mail addresses: niladri@iitg.ernet.in (N. Sett), ranbir@iitg.ernet.in (S. Ranbir Singh), sukumar@iitg.ernet.in (S. Nandi).

http://dx.doi.org/10.1016/j.neucom.2014.11.089
0925-2312/© 2015 Elsevier B.V. All rights reserved.

To study the influence deeper, we systematically explore it in three levels. First, plain weighting models are applied over datasets. On the first set of datasets, the weighted model consistently outperforms its unweighted counterparts, and on the second set of datasets, unweighted model consistently outperforms its weighted counterparts. For other datasets, their performances are only marginally different. Second, we modify the weighting models by applying tuning parameters. With a proper selection of tuning parameters, a significant boost in the performance of the weighted models is observed. After tuning, the weighted models perform better than its unweighted counterparts even for some of the second set of datasets, which has been observed otherwise before tuning. Third, from the experimental observations it is evident that an effect of the tie weight depends on the characteristics of the dataset. In the light of this, we present a neighborhood based analysis of datasets to find out the reason behind the diversified effect of the tie weight on the node proximity based link prediction methods. We further extend the analysis based on density of the neighborhood of participating nodes,[1] by introducing *odd ratio* over the node degree.[2] It is interesting to observe that for the participating nodes with low average odd ratio, weighted models are suitable, and for the nodes with high average odd ratio, unweighted methods are suitable. In short, we can summarize our contributions as follows:

- Propose min-flow and multiplicative weighting models.
- Investigate the effect of three weighting models; additive, min-flow and multiplicative on the prediction methods and datasets.
- Systematically study the effect of weight on prediction methods by introducing weighted links iteratively.
- Analyze the effect of two weight tuning methods and apply over RA.
- Present a node proximity based analysis of underlying graph to justify our results.
- Define degree odd-ratio and use it to propose a directive model for effective prediction.

The rest of the paper is organized as follows. Section 2 presents the existing node proximity based prediction methods and the proposed weighting models. The experimental datasets are discussed in Section 3. Sections 4–6 discuss different observations and analysis in details. Finally, Section 7 concludes the paper.

2. Prediction methods: weighted and unweighted

The prediction methods CN, JC, AA and RA explore the local proximity of two nodes to estimate a predicted score. A classical comparative study of various prediction methods (including the first three) has been reported by Liben-Nowell and Kleinberg in [2]. Later in 2009, the resource allocation measure has been introduced by Zhou et al. [7]. All these measures assign a positive score to a node pair, if and only if there is at least one 2-length path between the participating nodes, i.e., the participating nodes have at least one common neighbor. Among these four methods, RA is reported to perform better in several studies [7,11,13], and all these studies except [11] have considered only unweighted networks.

Study on the effect of tie weights over the local analysis based node proximity measures is still not explored much. The first such study has been presented by Murata et al. [10]. Authors have investigated the effect on three measures; CN, JC and AA using Yahoo! Chiebukuro social graph. Their results indicate a positive influence of tie weights on the link prediction. However in [11], the authors revisited the problem and observed conflicting results i.e., the performance of weighted measures of almost all proximity measures (CN, AA and RA) are worsen in all of the three datasets; USAir (US air transportation network), C.elegans (neural network of the nematode worms) and CGScience (co-authorship network of computational geometry). Lü et al. have further extended the study to investigate the role of weak ties and concluded that their results have been influenced by Granovetter's weak tie theory [8], i.e., weak ties play an important role in the information dissemination in social networks. In [12], the authors have not found significant improvement in performance, while experimenting with weighted co-authorship networks. However, the performance improved when they have applied supervised approach to the weighted measures. In another recent study [14], the authors have explored face to face interaction network among researchers and have observed that the weighted methods outperform their unweighted counterparts.

In all these studies, only an additive (linear summation) model has been used to incorporate weights. However, the additive model may not have equal effect on different prediction methods. Like existing studies, this paper also focuses on CN, JC, AA and RA, but investigates the responses of three different weighting models: (i) additive, (ii) min-flow and (iii) multiplicative.

2.1. Three weighting models

If x and y are the two participating nodes, the *additive* strength between x and y is bound by a common neighbor z, which is defined by a linear model $w(x,z)+w(z,y)$, where $w(-,-)$ is the symmetric edge weight connecting two nodes. If we assume that two nodes x and y have infinite supply of information through channels connecting them, $w(x,z)+w(z,y)$ represents the aggregate information received by node z from nodes x and y. The higher the volume of information z receives, the tighter is the bond that z holds between x and y. Fig. 1(a) shows the graphical representation of the additive model, where z acts as an information aggregator. Thickness of the edges represents the strength of the tie.

Unlike additive model, *min-flow* defines the bonding between x and y by the channel capacity, $\min(w(x,z),w(z,y))$, through z. Considering channels of different capacities, the information received by one node from another node is defined by the channel of lower capacity. Fig. 1(b) shows the graphical representation of min-flow measure, where z acts as a flow control node between x and y.

In *multiplicative* model, node z acts as a flow booster. The incoming flow is exaggerated by many folds defined by the outflow channel capacity i.e., $w(x,z)\times w(z,y)$. In the following subsections, we incorporate the above three weighting models with each of the prediction methods and define the weighted versions.

2.2. Prediction methods

In this section, we incorporate above three weighting models with each of the prediction methods and define their weighted estimates.

2.2.1. Common Neighbor (CN)

The idea behind the common neighbor index in a social network graph is that – if two actors (nodes) x and y have many friends in common, they are more likely to form a link in the future. If $\Gamma(x)$ and $\Gamma(y)$ denote the set of neighbors of x and y

[1] Nodes connected by strong ties are considered to belong to the same region or community and nodes connected by weak ties are considered to belong to different regions or communities.

[2] Ratio between unweighted and weighted.

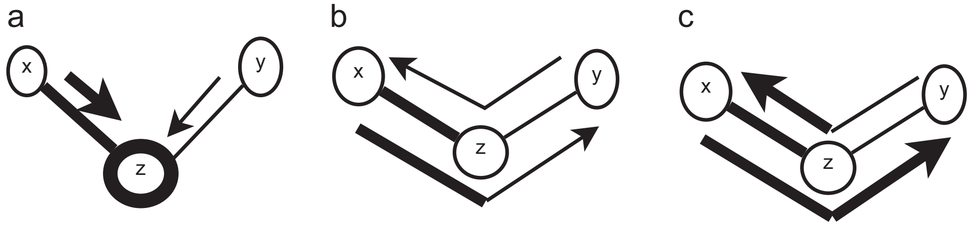


Fig. 1. Graphical representation of different weighting models.

respectively, the unweighted CN score is defined as below:

$$CN(x,y) = |\Gamma(x) \cap \Gamma(y)|$$

In [10], Murata et al. define the additive weighting model as below:

$$CN_A(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} (w(x,z) + w(z,y)).$$

In this, the binding strength between x and y is defined by the collective information received by all common neighbors. Similarly, we define the min-flow and multiplicative model of CN with the collective amount of information received by one node from another through all common neighbors, as follows:

$$CN_{MF}(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \min(w(x,z), w(z,y)),$$

where $\min(w(x,z), w(z,y))$ defines the channel capacity passing through z, and

$$CN_M(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} (w(x,z) \times w(z,y))$$

In multiplicative, z acts as a booster node. The incoming strength towards z is either boosted or filtered by the outgoing channel capacity.

2.2.2. Adamic/Adar (AA)

Adamic/Adar measure treats each common neighbor differently. Common neighbors, having less number of connections, contribute more into the score, as a node having less number of connections associates more closely with its existing connections than a node having large number of connections. The idea of AA has been borrowed from [1], where Adamic and Adar have proposed a similarity measure between two homepages. The unweighted Adamic/Adar score is defined as

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

In [10], Murata et al. define the additive weighted measure of AA as follows:

$$AA_A(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,y)}{\log(1 + s(z))},$$

where $s(x) = \sum_{z \in \Gamma(x)} w(x,z)$ is the additive strength or weighted degree of node x. Like CN, we define the min-flow and multiplicative model of AA as below.

$$AA_{MF}(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\min(w(x,z), w(z,y))}{\log(1 + s(z))},$$

and

$$AA_M(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) \times w(z,y)}{\log(1 + s(z))},$$

In all the equations of weighted versions of AA, one is added with $s(z)$ to avoid negative score.

2.2.3. Resource Allocation (RA)

Resource Allocation method is much like Adamic/Adar. The unweighted RA score is given by [7]

$$RA(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|}.$$

Lü et al. [11] have defined the additive weighted measure of RA as below:

$$RA_A(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,y)}{s(z)},$$

We introduce the min-flow and multiplicative models as

$$RA_{MF}(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\min(w(x,z), w(z,y))}{s(z)},$$

and

$$RA_M(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) \times w(z,y)}{s(z)},$$

2.2.4. Jaccard's Coefficient (JC)

The motivation of this measure is derived from a paper [15], written by Jaccard, way back in 1901. The score of unweighted JC is given by

$$JC(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}.$$

JC assumes that the strength of collaboration decreases with the increase in the number of associations with the two participating nodes. The additive, min-flow and multiplicative model of JC are formulated in our paper as follows:

$$JC_A(x,y) = \frac{\sum_{z \in \Gamma(x) \cap \Gamma(y)} (w(x,z) + w(z,y))}{s(x) + s(y) - \sum_{z \in \Gamma(x) \cap \Gamma(y)} \min(w(x,z), w(z,y))},$$

$$JC_{MF}(x,y) = \frac{\sum_{z \in \Gamma(x) \cap \Gamma(y)} \min(w(x,z), w(z,y))}{s(x) + s(y) - \sum_{z \in \Gamma(x) \cap \Gamma(y)} \min(w(x,z), w(z,y))},$$

and

$$JC_M(x,y) = \frac{\sum_{z \in \Gamma(x) \cap \Gamma(y)} (w(x,z) \times w(z,y))}{s(x) + s(y) - \sum_{z \in \Gamma(x) \cap \Gamma(y)} \min(w(x,z), w(z,y))}.$$

The denominator $s(x) + s(y) - \sum_{z \in \Gamma(x) \cap \Gamma(y)} \min(w(x,z), w(z,y))$ represents the weighted equivalent of $|\Gamma(x) \cup \Gamma(y)|$ as $|\Gamma(x) \cup \Gamma(y)| = |\Gamma(x)| + |\Gamma(y)| - |\Gamma(x) \cap \Gamma(y)|$. Here we consider min-flow weighted version for the subtraction.

3. Datasets

We consider eight datasets of diverse characteristics to study the effect of above three weighting models. A brief characteristics of the datasets are presented in Table 1. DBLP and Enron networks are constructed locally from the raw data that includes a timestamp. For other datasets, ten-cross validation is used.

Table 1
Characteristics of the datasets.

Datasets	#Nodes	#Edges	Avg CC	Avg degree
DBLP	339,223	969,287	0.643	5.715
Enron	76,548	297,224	0.153	7.766
Newman	16,264	47,594	0.562	5.853
OCLinks	1899	13,838	0.109	14.574
Openflights	2939	15,677	0.453	10.668
Astro	16,046	121,251	0.665	15.113
Hep-th	7610	15,751	0.486	4.140
NetScience	1461	2742	0.694	3.754

Considering the nature of the network, the datasets are divided into three groups. A brief discussion is presented below:

1. Collaboration Networks
   - DBLP[3]: It is a co-authorship network of computer scientists over 7 years (between 2001 and 2007). Co-authorship information during first 5 years is used to build the training graph, and testing edges are constructed from the co-authorship information during the last 2 years. Based on previous studies found in the literature, we create two datasets using two different tie weighting methods (1) traditional number of co-authorships and (2) Newman's method, where for each collaboration, $1/(n-1)$ is contributed to the link weight, n being the number of authors present in the collaboration [16]. We name them as Dblp-1 and Dblp-2 respectively.
   - Newman, Astro, Hep-th, and NetScience: Newman, Astro and Hep-th [16] are collaboration networks of Condensed matter, Astrophysics and High-energy theory of arXiv E-Print Archive between 1995 and 1999. NetScience [17] is a collaboration network of network scientists. Newman collaboration network dataset is available with both the weighting schemes used for DBLP, and we name them as Newman-1 and Newman-2, respectively. Astro, Hep-th, and NetScience use Newman's method as the tie weighting measure.
2. Communication Networks
   - Enron[4]: It is an e-mail network built over the time spanning January 1997–December 2002. Email network till November 2001 is considered as the training dataset and remaining as the test network. Tie weight is the number of emails exchanged between two individuals.
   - OCLinks: OCLinks [18] is a communication network dataset collected from a Facebook like social network. This network is built upon the message exchange among the users. The weight of a link indicates the number of messages exchanged between two users.
3. Other
   - Openflights[5]: This is a network of airports. The number of routes between two airports gives the weight of the link between them.

The above datasets are further grouped into four subsets based on the nature of the network and the type of edge weighting method, which are referred as `Col-1`: Dblp-1 and Newman-1; `Col-2`: Dblp-2, Newman-2, Astro, Hep-th and NetScience; `Com`: Enron and OCLinks; and `oth`: Openflights.

3.1. Evaluation

Class imbalance is one of the issues in the link prediction problem that affects the representation of the performance [19]. A popular method to address the class imbalance issue is to use the area under a ROC curve (AUC) [20]. ROC plots true positive rate vs. false positive rate for all thresholds. The AUC value is equivalent to the probability of a randomly chosen positive instance having higher prediction score than a randomly chosen negative instance [7,19,21]. Considering high imbalance between number of existing links (probe links that appear in the testing period i.e., positive instances) and non-existing links (links that do not appear i.e., the negative instances), AUC has become a common evaluation metric in recent studies [7,22,11,14]. This paper also uses AUC measure to evaluate the performance. The detailed method is as follows.

Let p be the number of probe links and n be the number of non-existing links; c_i and d_i denote the number of non-existing links that have less and equal score respectively, compared to probe link i. Then AUC estimation can be simplified as follows [13]:

$$AUC = \frac{\sum_i c_i + 0.5 \times \sum_i d_i}{n * p}$$

The number of non-existing links is extremely large compared to the number of existing links. Comparing the score of an existing test link with that of all non-existing links is computationally very expensive. In [13], this issue is addressed by random selection of reasonably large number of non-existing links. We adopt similar approach in this paper. However, selecting the appropriate number of non-existing links is an important issue. An inappropriate number may result in a biased score. In the experiments reported in this paper, we have decided upon the number of non-existing links after an analysis of AUC scores that do not differ up to the third precision.

4. Experimental observations

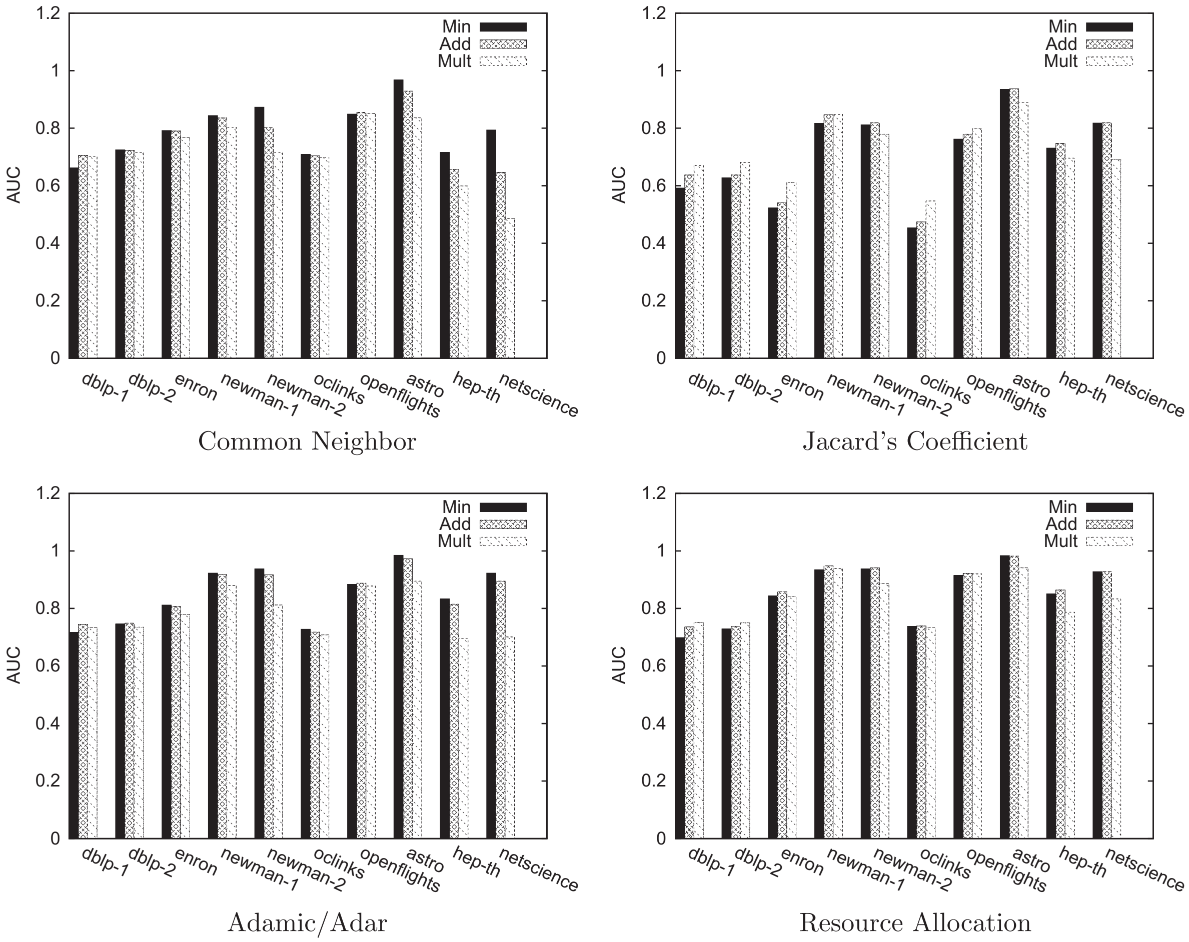
This section discusses the experimental responses of the edge weights on four local proximity based link prediction methods namely JC, CN, AA and RA. Previous studies [10–12,14,23] have discussed responses of the additive model. This paper introduces two other weighting models (as discussed in Section 2) and compare their responses upon a larger collection of datasets with the response of the methods reported in [10–12].

4.1. Response of different weighting models and networks

Fig. 2 compares the AUC scores of three weighting models namely min-flow, additive and multiplicative for CN, JC, AA and RA. There are 40 independent experimental cases (i.e., four local proximity based link prediction methods and 10 datasets), and following observations are evident.

- Characteristics of the relative response of the three weighting models vary from one dataset to another. Out of the 40 experimental cases, min-flow responds better than its counterparts in 42.5% of the cases, whereas additive and multiplicative in 37.5% and 17.5% of the cases respectively (see Table 2).
- For the CN and AA, min-flow outperforms its counterparts in 90% and 70% cases respectively. Although, for RA, additive responds better (in 60% of the cases), and for JC, there is a tie between multiplicative and additive (both 50% of the cases).
- From the comparison between the frequency based edge weight and the Newmans edge weight, min-flow weighting model responds positively on Newman's edge weight (i.e., min-flow responded better in 50% of the cases, additive in 35% and

[3] http://dblp.uni-trier.de/xml/
[4] http://www.cs.cmu.edu/~enron/
[5] Opsahl, T., 2011. Why Anchorage is not (that) important: Binary ties and Sample selection.



Common Neighbor

Jacard's Coefficient

Adamic/Adar

Resource Allocation

Fig. 2. Response of weighting models on prediction methods and datasets.

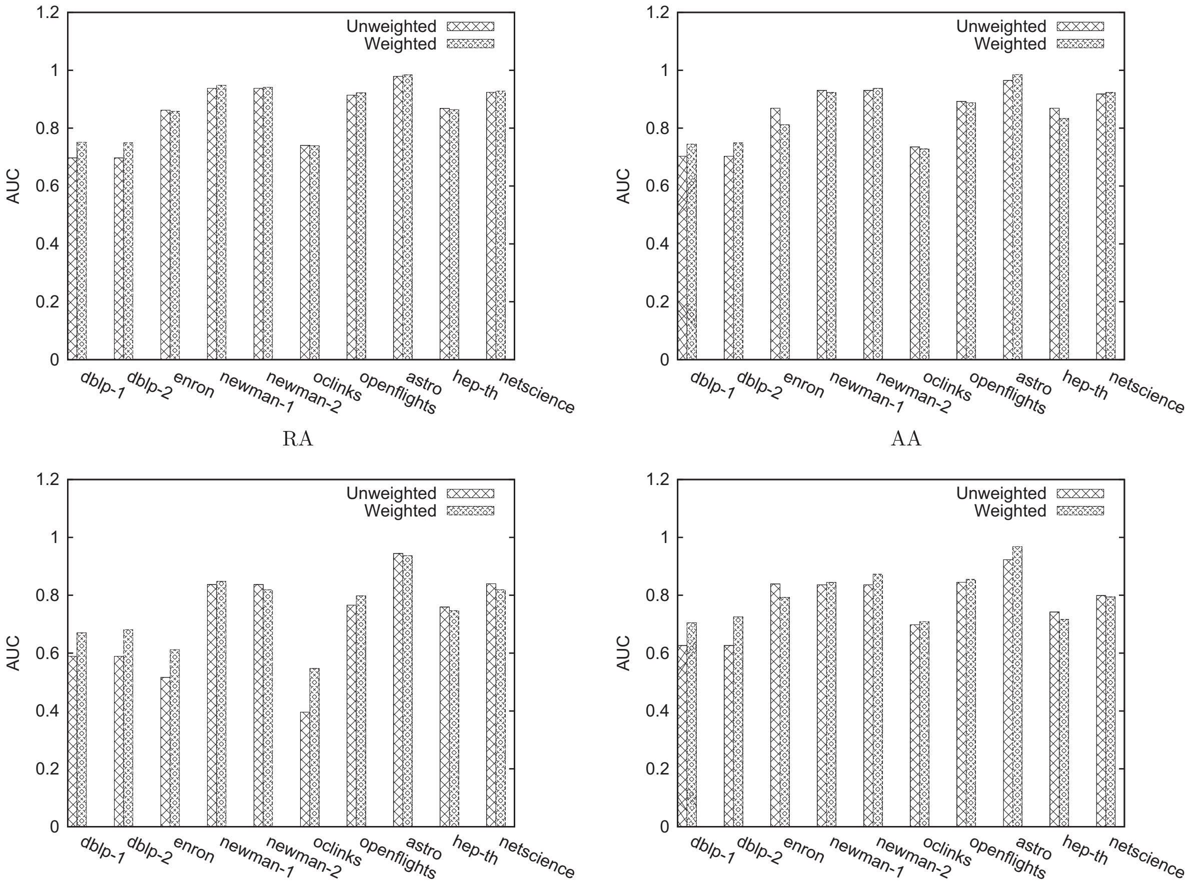
Table 2
Detailed experimental cases. (For every fraction, the denominator gives the total number of experimental cases and the numerator denotes the number of occasions where the particular method outperforms others for the particular dataset type. $x, y : z$ in the numerator means that in x number of cases the particular method outperforms others and in y number of cases a tie happens with weighting model z. % denotes the percentage of outperforming cases over all datasets.)

Method	Min-flow (f)				Additive (a)				Multiplicative (m)			
	Col-1	Col-2	Com	oth	Col-1	Col-2	Com	oth	Col-1	Col-2	Com	oth
CN	$\frac{1}{2}$ 90%	$\frac{5}{5}$	$\frac{2}{2}$	$\frac{1}{1}$	$\frac{1}{2}$ 10%	$\frac{0}{5}$	$\frac{0}{2}$	$\frac{0}{1}$	$\frac{0}{2}$ 0%	$\frac{0}{5}$	$\frac{0}{2}$	$\frac{0}{1}$
JC	$\frac{0}{2}$ 0%	$\frac{0}{5}$	$\frac{0}{2}$	$\frac{0}{1}$	$\frac{1}{2}$ 50%	$\frac{4}{5}$	$\frac{0}{2}$	$\frac{0}{2}$	$\frac{1}{2}$ 50%	$\frac{1}{5}$	$\frac{2}{2}$	$\frac{1}{1}$
AA	$\frac{1}{2}$ 70%	$\frac{4}{5}$	$\frac{2}{2}$	$\frac{0}{1}$	$\frac{1}{2}$ 30%	$\frac{1}{5}$	$\frac{0}{2}$	$\frac{1}{1}$	$\frac{0}{2}$ 0%	$\frac{0}{5}$	$\frac{0}{2}$	$\frac{0}{1}$
RA	$\frac{0}{2}$ 10%, 10% $:a$	$\frac{1,1:a}{5}$	$\frac{0}{2}$	$\frac{0}{1}$	$\frac{1}{2}$ 60%, 10% $:f$	$\frac{2,1:f}{5}$	$\frac{2}{2}$	$\frac{1}{1}$	$\frac{1}{2}$ 20%	$\frac{1}{5}$	$\frac{0}{2}$	$\frac{0}{1}$
Overall	42.5%, 2.5% $:a$				37.5%, 2.5% $:f$				17.5%			

multiplicative in 10% on `Col-2` datasets). However, additive performs best in the case of frequency based edge weight. Of all the cases over `Col-1`, `Com` and `oth` datasets, additive perform best in 40% of the cases, min-flow in 35% of the cases and multiplicative in 25%.

Above observations clearly indicate that the performance of weighting models depends on the characteristics of the dataset and the subjected prediction method. Therefore, subject to the dataset and prediction method, weighting model should be carefully chosen.

Fig. 3 further compares the performance of weighted and unweighted versions of CN, JC, AA and RA. For each prediction method and dataset pair, the best responding weighted model and its unweighted counterpart are selected and plotted in Fig. 3. Like in Fig. 2, the responses from the weighted and unweighted methods are evenly distributed throughout the experimental cases. It clearly shows that the response of edge weight depends on many factors,



RA AA

JC CN

Fig. 3. Comparing the AUC score of weighted and unweighted link prediction counterparts.

such as (i) type of the network, (ii) weighting model of the link prediction and (iii) the edge weight formation. It is important to study the influence of these factors on the prediction performance. The remaining part of the paper attempts to analyze the influence of each of these factors on the link prediction. We also analyze the effect of tuning the weighting models. In the next subsection, we investigate the effect of link weight over datasets of different nature.

4.1.1. *Effect of edge weight on collaboration network*

As mentioned in Section 3, collaboration networks are divided into two groups namely `Col-1` and `Col-2`, based on the way edge weights are defined. Referring to the plots of Dblp-1 and Dblp-2 datasets in Fig. 3, a clear case of favoring weighted methods is evident. For all the eight cases, weighted methods respond positively and there is no significant difference in performance between Dblp-1 and Dblp-2 over the methods.

Referring to the plots related to Newman-1 and Newman-2 datasets in Fig. 3, it is surprising to observe that there is no clear winner between unweighted and weighted methods. Though `Col-1` and `Col-2` are of similar nature (i.e., collaboration networks), the observations in Fig. 3 contradict each other in Newman dataset. It also indicates that the tie weighing method also influences the link prediction performance.

4.1.2. *Effect of edge weight on communication networks and Openflights*

Though the weighted methods outperform their unweighted counterpart for Openflights, unlike collaboration network, unweighted methods tend to perform better than weighted methods in communication networks (refer Fig. 3). The contradictory response of prediction methods in these datasets indicates the presence of highly centered nodes in the graph. In Enron and OCLinks datasets, few nodes have connectivity with a large fraction of the total nodes in the network. A large number of participating nodes are neighbors of these centered nodes, and the connecting edges to these centered nodes have high edge weights (see Section 5.1). Because of these edges, scores of undeserving node pairs are expected to be boosted inappropriately, resulting in lower AUC scores.

Even after tuning on weighting parameters, discussed in Section 4.3, the performance drops. Therefore, these two tuning methods are not suitable for the datasets having highly centered nodes. However, suitable study on tuning methods for the dataset having highly centered node is beyond the scope of this paper. Instead, we investigate density based enhancement (see Section 6), where performance of the weighted models is significantly enhanced.

4.2. *Effect of edge weights on prediction methods*

This section investigates the influence of edge weights from different perspectives by introducing edges in the network in an incremental fashion. In this experiment, the graph is initially assumed to be empty. Then the edges are introduced in increasing order of their edge weights. The edges are divided into 10 folds based on their weights. This experiment helps us to understand influence of the edge weights over the link prediction methods. The plots in Fig. 4 show an interesting characteristic.

- Prediction performance gradually increases with the increase in the tie weights for both JC and CN. However, plots quickly

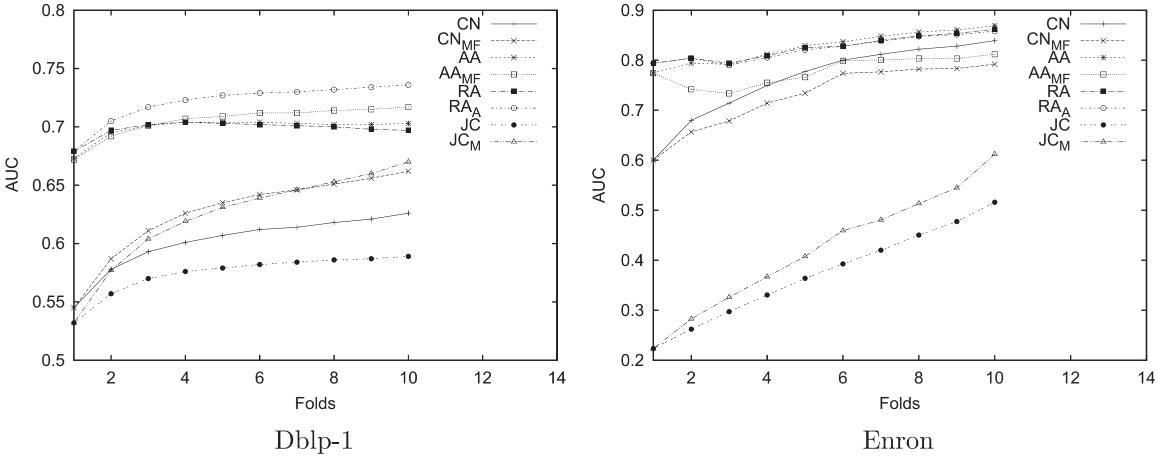


Fig. 4. Effect on different prediction methods after introducing edges in increasing order of their weights.

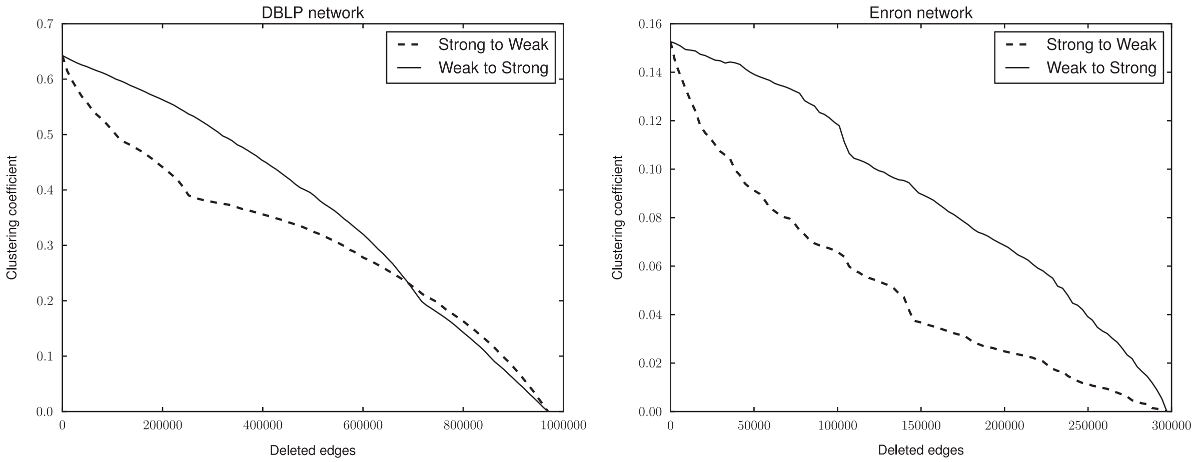


Fig. 5. Decrease of clustering coefficient by deleting links in increasing or decreasing order of their weights.

stabilize in the case of AA and RA. It indicates that AA and RA are more resistive towards the tie weights as compared to JC and CN. Hence, one can expect a larger effect of the link weight on JC and CN compared to AA and RA, which is depicted in Fig. 3 as well.

In Fig. 5, we further analyze the change in the clustering coefficient of two graphs while deleting the edges in an increasing or a decreasing order of the edge weights. The two graphs have a significant difference in characteristics – (i) the area between the two curves is small in DBLP and large in Enron; and more interestingly, (ii) the two plots intersect in the case of DBLP and does not intersect in the case of Enron. The rate at which the plot decreases, indicates the change in the local density surrounding the nodes. While deleting edges starting with stronger to weaker, DBLP gradually decreases its clustering coefficient, while for Enron, it exponentially decreases. It indicates that the cohesiveness of Enron is bounded by few central nodes with highly weighted edges, whereas, it is distributed uniformly in the case of DBLP. These observations provide an indication that the response of the tie weight on prediction methods depends on the characteristics of the underlying graph. A detailed study on the clustering coefficient is presented in Section 5.1.

4.3. Tuning the weighted models

This section investigates if the performance of the prediction method can be improved by tuning the weighting factor. It proposes two tuning methods namely; *scaling* and *linear sum*. These tuning methods are applied on RA. We have selected RA for this study, because RA consistently performs better than other prediction methods as seen in Fig. 6. Out of the 40 cases, RA performs better than others in 29 cases, and RA and AA jointly outperforms in 9 more cases. Further, it is observed in Table 2 that RA performs best with the additive model. This subsection focuses on RA using the additive model.

4.3.1. Scaling

This tuning method is motivated by three cases of additive model based weighted RA as shown in Fig. 7; the three cases of a subgraph between participating nodes x and y with their common neighbor z. The thicker line represents a higher weight. Additive model based weighted RA returns same score for all these cases because, there is no neighbor of z other than x and y. However, their response should be different. Considering the triadic closure property of the weak tie theory [8] i.e., *if edges* (x,z) *and* (y,z) *are connected by strong ties*, *the chances of connecting x and y by at least a weak tie is high*, the prediction score should be biased by the tie weight of the edges connecting the common neighbors. Accordingly, the tuning factor is introduced as follows:

$$RRA_s(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{(w(x,z)+w(z,y)).\Omega(x,y)}{s(z)}$$

where $\Omega(x,y)$ is the tuning factor. In this study, the tuning factor is defined as $\Omega(x,y) = \log(w(x,z)+w(z,y))$. It is similar to the *rich get*

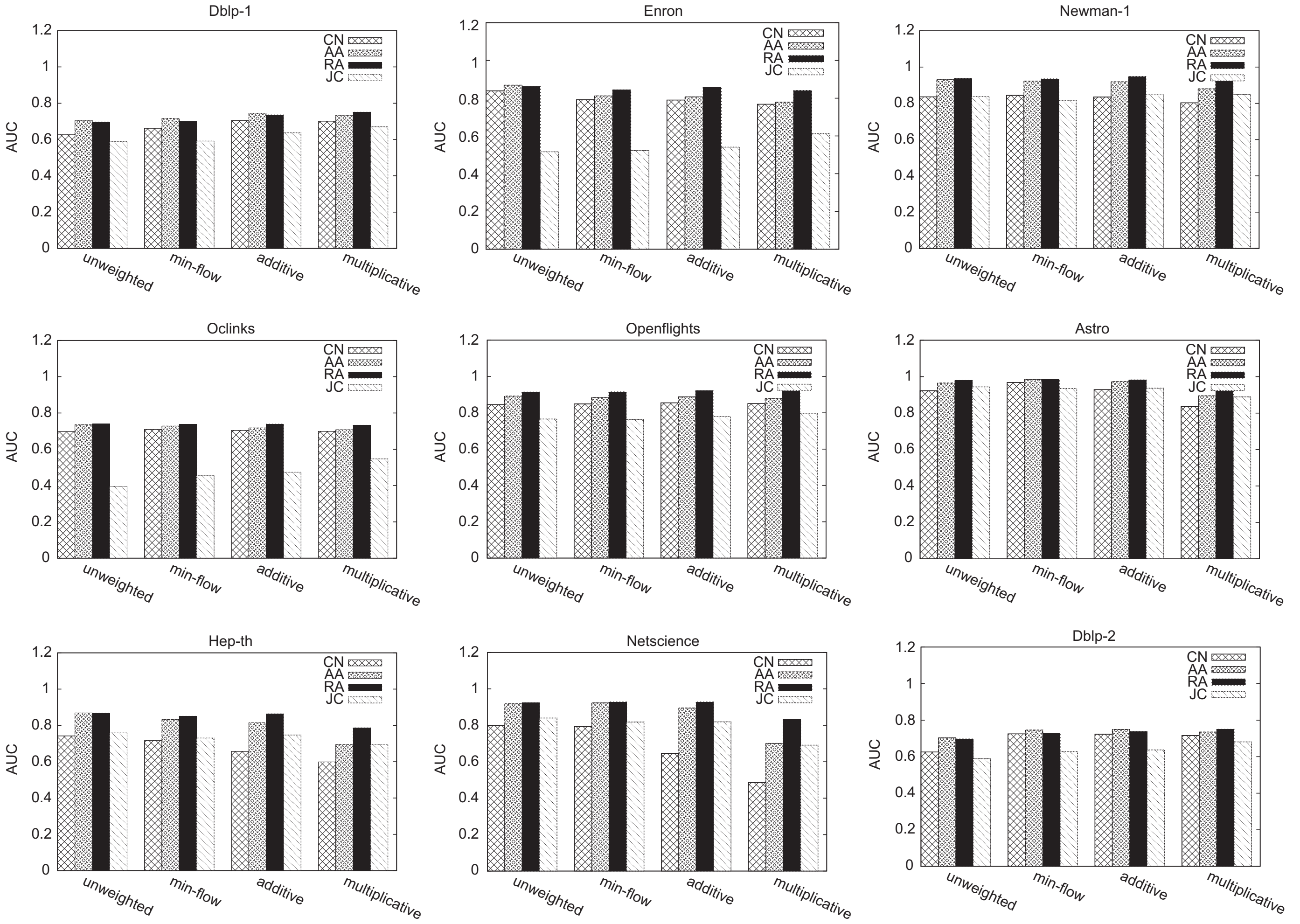


Fig. 6. Performance of prediction methods over datasets and weighting methods. Plots for Newman-2 are excluded from the graph to accommodate the plots in three rows.

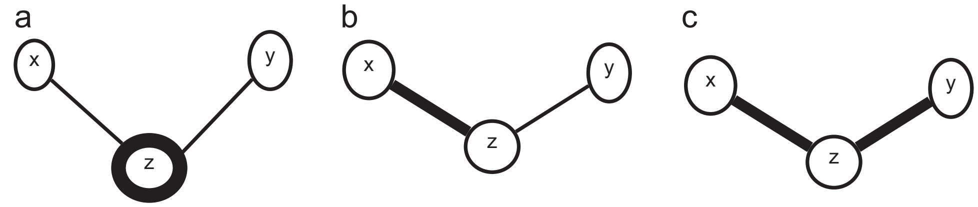


Fig. 7. Three cases of additive models of RA.

richer phenomenon. If the nodes x and y are connected by a neighboring node z with strong tie weight, then the score is boosted by a larger factor.

4.3.2. *Linear Sum*

In each of CN, AA and RA, the importance of $z \in \Gamma(x) \cap \Gamma(y)$ alone is considered into account. The relative importance of the participating nodes is entirely ignored. However, participating nodes, which are having relatively higher amount of interaction with common neighbors, are expected to reflect stronger bonding. RA is regularized using linear summation as follows:

$$RRA_l(x,y) = \alpha \cdot \Omega(x,y) + (1-\alpha) RA_A(x,y)$$

where $0 \le \alpha \le 1$ and $\Omega(x,y)$ is the regularization factor. In this study, we use $\frac{1}{2} \times \sum_{z \in \Gamma(x) \cap \Gamma(y)} (w(x,z)/s(x) + w(y,z)/s(y))$ as $\Omega(x,y)$.

4.3.3. *Effect of weight after tuning*

Table 3 compares the performance of the weighted RA with tuning and its counterparts; unweighted RA and additive weighting model. It clearly shows that the tuning methods affect the datasets differently. Regularization using linear weighted sum (denoted by RRA_l) enhances the performance significantly over all collaboration networks for both the frequency based and Newman's edge weighting measure [16]. However, scaling enhances only for the networks whose edges are weighted using frequency. Fig. 8 compares the performance of unweighted RA and best weighted RA. With a right choice of weighting and tuning methods, edge weights can indeed enhance the link prediction performance over the collaborative networks.

Surprisingly, there is no influence of the above tuning methods on Openflights datasets, and there is negative influence of the tuning methods on communication networks. This contradicting

Table 3
Effect of regularized RA.

Dataset	RA	RA_A	RRA_S	RRA_l	Dominant
Dblp-1	0.697	0.736	0.747	0.736	RRA_s
Dblp-2	0.697	0.738	0.747	0.738	RRA_s
Enron	0.862	0.858	0.853	0.682	RA
Newman-1	0.938	0.948	0.944	0.958	RRA_l
Newman-2	0.938	0.941	0.934	0.952	RRA_l
OCLinks	0.740	0.739	0.737	0.683	RA
Openflights	0.914	0.922	0.921	0.921	RA_A
Astro	0.979	0.982	0.939	0.985	RRA_l
Hep-th	0.868	0.864	0.806	0.876	RRA_l
NetScience	0.924	0.928	0.844	0.937	RRA_l

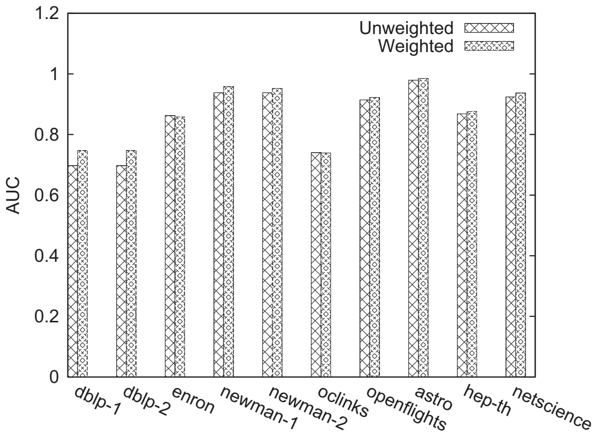


Fig. 8. Comparison between AUC scores of unweighted RA and best weighted measure among all variants of weighted RA.

behavior may be due to the presence of highly centered nodes in communication and Openflights networks. For such datasets, further investigation is needed on tuning methods, which is beyond the scope of this paper.

5. Node proximity based analysis of DBLP and Enron datasets

From the empirical study presented in the previous sections, it is observed that the effect of incorporating weight to the common neighbor based link prediction methods is not consistent over different networks. In the previous sections, analysis has been limited to applying different weighting models to local proximity based link prediction methods. Further, with suitable weight tuning methods, it is also observed that prediction methods can use the weight effectively. With the tuning methods introduced in Section 4.3, the prediction performance of all the collaboration networks improves with the weight factor. However, for communication networks, the response is observed otherwise. Such observations raise two questions – (1) which property of the network governs the effect of the link weight on the common neighbor based link prediction methods, and (2) given a network graph, can we effectively decide beforehand whether to incorporate weight or not? This section addresses these questions.

One such study has been presented in [11]. It has been reported that incorporating link weights with the local proximity indices worsen the performance of link prediction. They have added an exponent to the link weights while calculating the score and have discovered that for most of the cases, the prediction performance improves for negative values of the exponent. They have concluded that their result has been supported by Granovetter's *weak-tie theory* [8] and have done a subgraph analysis on the network graphs. Though they have claimed it to be *Motif analysis*, there is no clear justification of considering all possible types of triad connecting weak and strong ties as Motif [24].

Formation of new links in a social network is governed by the *triadic closure property* [8] (also referred as triangle closing), which states that, if two nodes have a common friend, then they are likely to be friends. Its weighted counterpart, the *strong triadic closure property* says that probability of two nodes being friends increases with the weight of the links that connect them with a common friend [8,25]. According to the subgraph analysis presented in [11], violation of the *strong triadic closure property* is clear. Their conclusion suggests that weak ties play major role in triangle closing. In this paper, clustering co-efficient of a node, which also relies on the *triadic closure property*, is analyzed to investigate the effect of weights on the link prediction measures. Dblp-1 and Enron graphs are chosen for analysis as the former one is a collaboration network, and the later is a communication network. These networks are considerably larger compared to others, and time-stamp information is also available for both, that allows us to test with true future links.

5.1. Analyzing clustering co-efficient (CC)

Clustering co-efficient (CC) [26] of a node x in a network can be given by $2|t(x)|/\Gamma(x)(\Gamma(x)-1)$, where $t(x)$ is the set of triangles formed by the node x and its neighbors. It reflects the density in the neighborhood of a particular node. High CC of a node indicates that the triadic closure property holds strong in its neighborhood. Average CC of a network, denoted as C, is defined as the clustering coefficient, averaged over all the nodes of that network.

Barrat et al. [27] have proposed a version of weighted CC that can be defined as (for undirected graph):

$$CC^w(x) = \frac{1}{s(x)(\Gamma(x)-1)} \sum_{\{x,y,z\} \in t(x)} (w(x,y)+w(x,z))$$

This measure not only quantifies the local cohesiveness in terms of triangle closure, but also takes the link weights into account to calculate the likelihood of forming triangles by a node with its neighborhood. The average weighted clustering co-efficient of a network C^w can be calculated in the same way as that of its unweighted version. The value of both C and C^w ranges between 0 and 1.

By the argument presented in [27,28], if $C^w > C$ holds in a particular network, it indicates that the triangle closing is governed by the stronger links, and the weaker ones dominate in the case of $C^w < C$. As per our result of link prediction on Dblp-1 and Enron graphs, this argument directs us to think that the first case holds in Dblp-1, and Enron falls in the second case. Surprisingly, both the networks show the same pattern $C^w > C$. In both the cases, the higher weights play more significant role than the lower weights in the phenomenon of triangle closing, and both the datasets support *strong triadic closure property*.

To further analyze these two graphs in terms of CC, inspired by [27], we compare $C(\Gamma)$ and $C^w(\Gamma)$, where $C(\Gamma)$ is the average CC of the nodes having degree Γ and $C^w(\Gamma)$ is its weighted counterpart. Fig. 9 shows the comparison of the spectrum of $C(\Gamma)$ and $C^w(\Gamma)$ for the two networks. For both networks, $C(\Gamma)$ and $C^w(\Gamma)$ decay as Γ increases. It supports the fact that the hub nodes with high degree, have lower CC as they work as bridges between different groups. However, the ratio between $C^w(\Gamma)$ and $C(\Gamma)$ is much higher in Enron network than Dblp-1, particularly for hub nodes. This is due to the fact that the concepts of hub nodes in Dblp-1 and Enron networks are different. In Dblp-1, the researchers, who work in interdisciplinary research areas and have a large amount of contacts spread throughout different geographical areas, act as

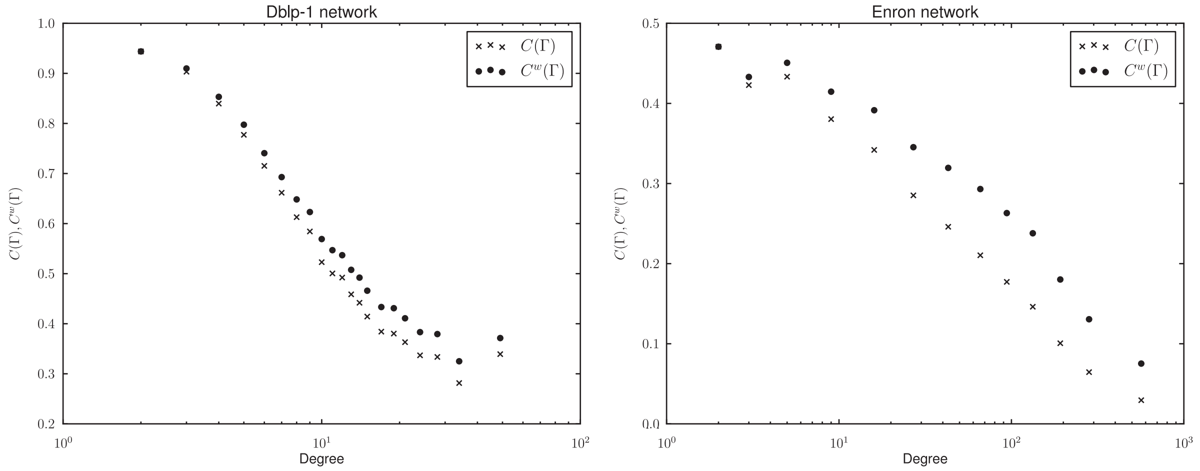


Fig. 9. Spectrum of $C(\Gamma)$ and $C^w(\Gamma)$ of Dblp-1 and Enron.

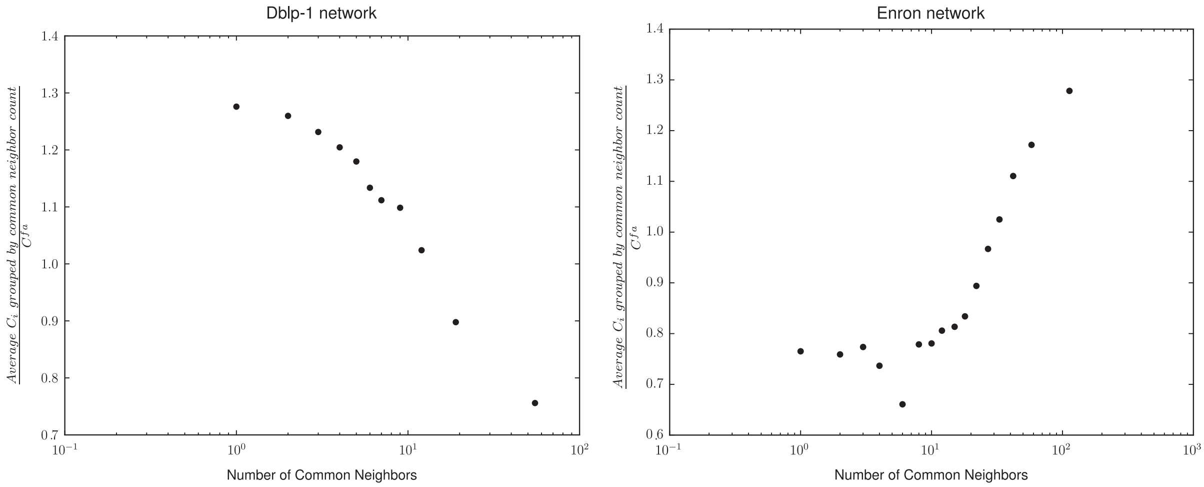


Fig. 10. Plots showing the average C_i, grouped by the number of common neighbors for Dblp-1 and Enron.

hubs. The hub nodes may represent experienced and prominent researchers. Furthermore, publishing a research paper is a slow process. Hence, the average weight of the edges connecting the hub nodes do not grow fast. On the contrary, as Enron is an e-mail network among the employees of an organization, the managers of the organization represent the hub nodes; they communicate with other employees very frequently. This keeps the ratio of $C^w(\Gamma)$ and $C(\Gamma)$ very high for Enron network.

Very high amount of weight surrounds the hub nodes in Enron network. The average degree of the hub nodes in Enron network is also very high compared to Dblp-1. This may cause the conflicting effect of edge weights in the link prediction performance on these datasets.

5.2. *The proposed measure*

From the findings in Section 5.1, it is evident that CC is not always able to answer the second question raised in the first paragraph of Section 5. CC measures the local cohesiveness of a particular node. However, all the measures presented in Section 2 rely on the number of common neighbors that a node pair have. Moreover, unlike average CC, which covers all the nodes of the network, the measures apply only on the node pairs that have at least one common neighbor. This motivates us to propose the following measure.

Table 4
AUC comparison between variants of RA for test edges in low and high region using degree odd ratio. An entry within parentheses shows percentage of either increase or decrease. The split in this table is 50–50%.

Region	Variant	Dblp-1	Dblp-2	Enron
High	RA	0.711(2.0)	0.697(0.1)	0.832(− 3.4)
	RA_A	0.710(− 3.5)	0.705(− 4.5)	0.818(− 4.6)
	RA_M	0.693(− 7.7)	0.674(− 10.5)	0.772(− 8.2)
Low	RA	0.682(− 2.2)	0.695(− 0.1)	0.890(3.3)
	RA_A	0.760(3.2)	0.771(4.5)	0.899(4.7)
	RA_M	0.807(7.4)	0.833(10.6)	0.910(8.2)

From the training graphs, we identify the edges for which the end nodes have at least one common neighbor. For each of such node pair (x_i, y_i), connected by edge e_i, we calculate the average of the edge weights associated with the common neighbors by

$$C_i = \frac{1}{K_i} \sum_{z \in \Gamma(x_i) \cap \Gamma(y_i)} (w(x_i, z) + w(z, y_i))$$

where $K_i = 2|\Gamma(x_i) \cap \Gamma(y_i)|$. This directly resembles with $CN_A(x, y)$, defined in Section 2.2. Next, we take the average of C_i's by $C^{ga} = N^{-1} \sum_{e_i} C_i$, where N is the number of edges having at least one common neighbor. We refer it as the *group average*.

We further take an average over all such edge weights by

$$C^{fa} = \frac{1}{K}\sum_i \sum_{z \in \Gamma(x_i) \cap \Gamma(y_i)} (w(x_i, z) + w(z, y_i))$$

where $K = \sum_i K_i$, and call it the *flat average*.

A very interesting observation follows when we calculate C^{ga} and C^{fa} for Dblp-1 and Enron networks respectively. $C^{ga} > C^{fa}$ holds for Dblp-1 and $C^{ga} < C^{fa}$ holds for Enron. This finding inspires us to inspect the relationship between C^{ga} and C^{fa} from the node proximity characteristics of a graph. The difference between C^{ga} and C^{fa} can be given by

$$C^{ga} - C^{fa} = \sum_i \left(\frac{1}{N} - \frac{K_i}{K}\right) \times C_i$$

As all the link weights considered here are of positive value, a particular node pair (x_i, y_i) contributes some positive value to the summation if $K_i < K/N$ holds for that particular pair. Therefore, if higher weights are concentrated in the edges of the common neighbors connecting the node pairs (x_i, y_i), which have lesser number of common neighbors, $C^{ga} - C^{fa}$ will give a positive value. Exactly similar scenario can be observed in Dblp-1 and Enron networks.

Fig. 10 shows the distribution of C_i, averaged over the number of common neighbors and normalized by C^{fa}, for the two networks. For Dblp-1, it clearly shows that the averaged C_i is high for the node pairs, whose number of common neighbors are small, and low when a number of common neighbors are large. Enron shows exactly the opposite characteristics.

From the analysis of C^{ga} and C^{fa}, we can loosely infer that if $C^{ga} > C^{fa}$ holds for a particular network, the weighted methods are likely to perform better than its unweighted counterpart and vice versa. Observation from Fig. 10 also supports the observation on the hub nodes presented in the previous subsection. For both the networks, it is observed that the number of common neighbors of two nodes increases with the average degree of the two nodes. So, in Enron network, when two manager nodes are linked and they have large number of common neighbors, as large weights are concentrated in their links, their C_i gives higher value.

6. Effect of localized weight distribution

One of the drawbacks of local node proximity based link prediction methods is its dependency on local density. In the previous section, we observed that for both Dblp-1 and Enron graphs, triangle closing is governed by stronger links. In [8], the author has said that close knit community is built by stronger ties and weak ties form *bridges* between communities. Motivated by these observations, we further investigate the effect of weighted and unweighted methods by distributing the participating nodes at different odd ratio. We define the odd ratio between unweighted and weighted as follows:

$$\text{Odd} = \frac{\text{unweighted score}}{\text{weighted score}}$$

We partition the test edges into two disjoint sets; *low* and *high*. We put an edge into `low` set if the average odd ratio between the two participating nodes is less than a threshold value. In this study, we sort the edges by the average odd ratio of participating nodes and divide the edge set into two at different split points i.e., 25–75%, 50–50% and 75–25%. For a 25–75% split, top 25% fall into `high` set and rest 75% in `low`. As mentioned earlier, RA_A performs best in most of the datasets and therefore we restrict our odd-ratio study only on RA_A.

In this section, we focus on the odd ratio of unweighted and weighted degree. The value of odd ratio of a node reflects the relative distribution of weights among its neighbors. Very high value of odd ratio means that the neighboring nodes are likely to be connected by weak ties, and very low values of odd ratio means that the neighboring nodes are likely to be connected by strong ties. If both the participating nodes have very low odd ratio value, the chances that both the participating nodes belonging to the same dense region or cluster is high. It is expected to achieve higher prediction accuracy compared to its counterparts in the high odd-ratio region. Table 4 clearly shows that the estimates in the low region outperforms the estimates in the high regions. Splitting the dataset into the high and low regions is like setting an experimental constraint to remove noisy links. With 50–50% splits, we are able to achieve an improvement of AUC up to 10% on the weighted measures and up to 2% on the unweighted measures.

Assuming that the network satisfies Granovetter's theory on tie strength, it is expected that the unweighted estimates in the high region decrease their performances, and the weighted estimates in the low region increase their performances as we increase the range split (say from 25–75% to 50–50% and then 75–25%). With small split range, the chances of nodes falling into different clusters in low region are high, and hence low region is expected to have noisy edges resulting in lower performance of weighted estimates. As we increase the split range, the amount of noisy edges gets decreased, resulting in the increase in the prediction performance of weighted estimates. It is also true for unweighted estimates that the high region at the large split range has higher number of noisy edges as compared to small split range. Hence, it results in decreasing prediction performance of unweighted estimates in the high region as we increase the split range. Fig. 11 clearly shows that the above expected

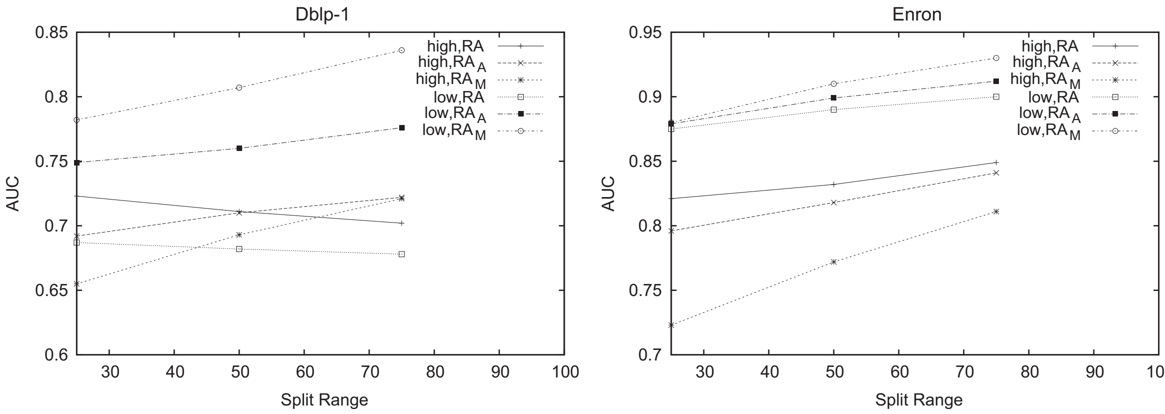


Fig. 11. Performance at different split range.

output is satisfied for DBLP dataset. In the case of Enron, with the increase in split range, the performances of all estimators are also increased. However, the rate (slope of the plot) at which unweighted estimate increases its performance is smaller than its weighted counterpart.

It can also be noted from Fig. 11 that the weighted estimates of the low region in all split ranges outperform its unweighted counterparts including Enron dataset. Therefore, with appropriate weighting model and appropriate tuning factors, weighted prediction method can perform better than its unweighted counterpart.

7. Conclusion

This paper presented an empirical analysis on the effect of the tie weight on four node proximity based link prediction methods: CN, AA, RA and JC. Two weighting models namely, min-flow and multiplicative are also introduced in this paper, both of which outperform the traditional additive model in a significant number of dataset-proximity measure pairs. The empirical results have shown a diverse effect of the tie weight on datasets and proximity measures. In few datasets, which are negatively affected by the tie weight, the weighted methods have been observed to perform positively when regularization is applied. However, the proposed regularization methods are not effective in few other datasets.

To understand the reason behind the conflicting responses of tie weight over different datasets, this paper has further exploited the characteristics of the underlying social network graphs. From this analysis, it is observed that the number of hub nodes in the network and their weighted strength influence the performance of the weighted node proximity based link prediction methods on that network. Further, the degree of odd-ratio analysis over the datasets has shown that unweighted models are effective for inter-cluster link prediction and weighted models are effective for intra-cluster link prediction.

References

[1] L.A. Adamic, E. Adar, Friends and neighbors on the web, Social Netw. 25 (3) (2003) 211–230.

[2] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, in: Conference on Information and Knowledge Management (CIKM'03), 2003, pp. 556–559.

[3] H. Kashima, N. Abe, A parameterized probabilistic model of network evolution for supervised link prediction, in: Proceedings of the Sixth International Conference on Data Mining, ICDM '06, IEEE Computer Society, Washington, DC, USA, 2006, pp. 340–349. http://dx.doi.org/10.1109/ICDM.2006.8.

[4] G. Salton, Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[5] Z. Huang, X. Li, H. Chen, Link prediction approach to collaborative filtering, in: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, ACM, 2005, New York, NY, USA, pp. 141–142.

[6] C. Wang, V. Satuluri, S. Parthasarathy, Local probabilistic models for link prediction, in: 2007 Seventh IEEE International Conference on Data Mining, ICDM 2007, IEEE, 2007, pp. 322–331.

[7] T. Zhou, L. Lü, Y.-C. Zhang, Predicting missing links via local information, Eur. Phys. J. B: Condens. Matter Complex Syst. 71 (4) (2009) 623–630 http://dx.doi.org/10.1140/epjb/e2009-00335-8.

[8] M. Granovetter, The strength of weak ties, Am. J. Sociol. 78 (6) (1973) 1360–1380.

[9] G.-J. Qi, C.C. Aggarwal, T. Huang, Community detection with edge content in social media networks, in: 2012 IEEE 28th International Conference on Data Engineering (ICDE), IEEE, 2012, pp. 534–545.

[10] T. Murata, S. Moriyasu, Link prediction of social networks based on weighted proximity measures, in: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, WI '07, IEEE Computer Society, Washington, DC, USA, 2007, pp. 85–88. http://dx.doi.org/10.1109/WI.2007.71.

[11] L. Lü, T. Zhou, Link prediction in weighted networks: the role of weak ties, EPL (Europhys. Lett.) 89 (2010) 18001 http://scholar.google.com/scholar.bib?q=info:KJLELzbFRhoJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,5&as_vis=1&scfhb=1&ct=citation&cd=0.

[12] H.R. de Sa, R.B. C. Prudêncio, Supervised link prediction in weighted networks, in: The 2011 International Joint Conference on Neural Networks, 2011, pp. 2281–2288.

[13] L. Lü, T. Zhou, Link prediction in complex networks: a survey, Physica A: Stat. Mech. Appl. 390 (6) (2011) 1150–1170.

[14] C. Scholz, M. Atzmueller, G. Stumme, On the predictability of human contacts: influence factors and the strength of stronger ties, in: Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, SOCIALCOM-PASSAT '12, IEEE Computer Society, Washington, DC, USA, 2012, pp. 312–321. http://dx.doi.org/10.1109/SocialCom-PASSAT.2012.49.

[15] P. Jaccard, Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines, Bull. Soc. Vaud. Sci. Nat. 37 (1901) 241–272.

[16] M.E.J. Newman, Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality, Phys. Rev. E 64 (2001) 016132.

[17] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (physics/0605087) (2006) 36104 21 p.

[18] T. Opsahl, P. Panzarasa, Clustering in weighted networks, Social Netw. 31 (2) (2009) 155–163.

[19] R. Lichtnwalter, N.V. Chawla, Link prediction: fair and effective evaluation, in: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE Computer Society, Los Alamitos, CA, USA, 2012, pp. 376–383.

[20] J.A. Hanley, B.J. Mcneil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36.

[21] A. Clauset, C. Moore, M.E. Newman, Hierarchical structure and the prediction of missing links in networks, Nature 453 (7191) (2008) 98–101.

[22] R.N. Lichtenwalter, J.T. Lussier, N.V. Chawla, New perspectives and methods in link prediction, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2010, pp. 243–252. http://dx.doi.org/10.1145/1835804.1835837.

[23] T. Herman, M. Monsalve, S. Pemmaraju, P. Polgreen, A.M. Segre, D. Sharma, G. Thomas, Inferring realistic intra-hospital contact networks using link prediction and computer logins, in: Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, IEEE Computer Society, Washington, DC, USA, 2012, pp. 572–578. http://dx.doi.org/10.1109/SocialCom-PASSAT.2012.113.

[24] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: simple building blocks of complex networks, Science 298 (5594) (2002) 824–827.

[25] M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a feather: homophily in social networks, Annu. Rev. Sociol. (2001) 415–444.

[26] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (6684) (1998) 440–442.

[27] A. Barrat, M. Barthelemy, R. Pastor-Satorras, A. Vespignani, The architecture of complex weighted networks, Proc. Natl. Acad. Sci. USA 101 (11) (2004) 3747–3752.

[28] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, J. Kertesz, Generalizations of the clustering coefficient to weighted complex networks, Phys. Rev. E 75 (2) (2007) 027105.

Niladri Sett received B.E. and M.Tech. degree from NIT Durgapur, India, in 2005 and 2009 respectively. Currently, he is working towards Ph.D. degree in the department of computer Science and Engineering, IIT Guwahati, Guwahati, India. His current research interest includes Social Network Analysis, Data Mining and Graph Theory.

Sanasam Ranbir Singh is an assistant professor in the Department of Computer Science and Engineering at Indian Institute of Technology Guwahati, India. He received a Ph.D. and a master degree in computer science and engineering from Indian Institute of Technology Madras, India. His research areas include Complex network analysis, information retrieval and machine learning. He is a member of IEEE Computer Society.

Sukumar Nandi received B.Sc. (Physics), B.Tech. and M. Tech. from Calcutta University. He received his Ph.D. from Indian Institute of Technology Kharagpur. Presently, he is a Professor in the Department of Computer Science and Engineering at Indian Institute of Technology, Guwahati. He was in School of Computer Engineering, Nanyang Technological University, Singapore as a Visiting Senior Fellow. He is the co-author of the book entitled "Theory and Applications of Cellular Automata", published by IEEE Computer Society. He has published over 250 Journal/Conference papers. His research interests include traffic engineering, wireless networks and information security. He is a Fellow of the Institution of Engineers (India), a Fellow of the IETE, a Senior member IEEE and a Senior member ACM.