Double Bounded Rough Set, Tension Measure and Social Link Prediction

Suman Kundu and Sankar K. Pal *Life Fellow, IEEE*

Abstract—The paper describes a new approach of viewing a social relation as a string with various forces acting on it. Accordingly, a tension measure for a relation is defined. Various component forces of the tension measure are identified based on the structural information of the network. A new variant of rough set, namely, double bounded rough set is developed in order to define these forces mathematically. It is revealed experimentally with synthetic and real-world data that positive and negative tension characterize, relatively, the presence and absence of a physical link between two nodes. An algorithm based on tension measure is proposed for link prediction. Superiority of the algorithm is demonstrated on nine real-world networks which include four temporal networks. The source code for calculating tension measure and link prediction algorithm is publicly available at https://gitlab.com/suman5/social-tension-measure.

Index Terms—Social network, rough set, link prediction, granular network, relational data, tension of relations, network evolution.

1 Introduction

A society is a group of people formed due to interactions among themselves. The word "society" came from a Latin word "societas" which was derived from the noun socius, used for the bond and interaction between friends and parties. Human beings are socialogical creatures and they have been living in a society for ages. With the development of science and technology the interactions among them have taken a new form. Along with the existing physical communication, they now form virtual connections with their peers. These virtual connections even grow beyond one's geographical boundaries. Facebook, Twitter, Flickr, Whatsapp etc. are some of the popular social networking apps which provide platforms for such long distance interactions among people.

Research shows that, unlike their physical counter parts, online social networks rapidly change over time. New links appear between actors and old edges become dormant over time. These very properties of network provide a new area of research and is referred as network evolution. Several researchers have studied its different aspects. These include empirical analysis of an evolving social network [1], actor-oriented models for co-evolving social networks and individual behaviors [2], time-aware link prediction in evolving networks [3], emergence of segregation in evolving social networks [4] and user preferences dynamics on evolving social networks [5], just to name a few.

A deep psychology works behind any relationship in a society and the same is also true for online social networks. A person usually gets motivated by the activities within its peers (neighbors). Understanding how an individual relationship works may help understanding the dynamics of the evolving network. In the present study we focused on an individual relationship and its local neighborhood to identify the different forces acting between the participating nodes of the said relationship. It is found that there exist at least four forces working on each individual relationship. Of them, two are positive and two are negative on the relations. The present paper proposes a tension measure on a social relation based on these forces. Tension force, in Physics, describes the pulling force transmitted axially by means of a string, cable, chain, or similar one-dimensional continuous object. A social relation may be viewed as a string connecting two nodes. The tension force on a social relationship, which is a novel concept, is the cumulative force which pulls the individual participants towards each other. A negative value of tension on a relation, on the other hand, indicates that the relationship is slack (i.e., inactive or no physical link between the participants in concern).

In order to find the value of the tension and the elementary forces constituting it, one needs to analyze the neighborhood of a relationship. Further, social networks show complex overlapping-neighborhood structures which are often indistinguisable, depicting granular structure [6], [7] (A granule is a clump of objects or points in the universe of discourse, drawn together, for example, by indistinguishability, similarity, proximity or functionality [8]). In such a system, the neighborhood of a relationship constitutes the granules which are in relation with either or both the participating nodes. Pawlak's theory of rough sets [9] provides a well known technology to extract knowledge from such indiscernible (ill-defined) information in granulated domain. Rough set theoretic approach is based on the principles of granular approximation of a set from its inner and outer sides concerning the belonging of granules to it. The theory hinges on the concept of two bounds of a set, namely, lower approximation and upper approximation which represent the sense of "definitely belonging", and "definitely and possibly belonging" of granules. Accordingly, the lower and (upper - lower) approximate regions

- *Suman Kundu is with the Center for Soft Computing Research, Indian Statistical Institute, Kolkata - 700108; Department of Computational Intelligence, Wroclaw University of Science and Technology, Wroclaw - 50-370; and ZINFI Software Systems Pvt. Ltd., Kolkata - 700091. (E-mail: suman@sumankundu.info)*
- *Professor Sankar K. Pal is with the Center for Soft Computing Research, Indian Statistical Institute, Kolkata - 700108. (Email: sankar@isical.ac.in)*

characterize two regions of a set or cluster, viz, the core (definite) and boundary (ambiguous) regions respectively. However, in social networks the neighborhood set of a relation consists of three types of elements, namely, granules which are related to both the participating nodes, and granules related to either of the nodes. These characteristics create three different types of regions in the neighborhood set. Granules which are related to both the participants constitute the core region of the neighborhood set, whereas, those related to either of the participating nodes constitute two types of boundary regions of the neighborhood set. Since we have two boundary regions, unlike Pawlak's rough set with single boundary region, we describe here a new variant of rough set, called *double bounded rough set* to deal with the situation.

Experiments with different synthetic and real world social networks reveal that the tension measure is positive for above 90% of linked pairs, and zero or negative for above 85% non-linked pairs for all the data sets. This shows a strong correlation between the tension measure and physical edges in the network. Because of this characteristic, the tension value is used for link prediction problem. Experiments have been conducted over nine real world network data. It is shown that our method is superior to the best known similarity based algorithms.

In summary, the contributions of this investigation are:

1) A new variant of rough set, namely, double bounded rough set has been introduced which can express the imperfect knowledge with relational data in granular framework.
2) The concept of viewing a social relation in terms of a string with different forces acting on it is unique. A tension measure depicting the cumulative force which pulls the nodes closer along the string has been provided.
3) The algorithm based on the tension measure to predict future links is new. This measure can also predict the removal of link in future.

The rest of the paper is organized as follows. The problem statement is presented in Section 2. In Section 3 we define the novel double bounded rough set. Different forces and the tension measure are described in Section 4. A solution of link prediction based on the tension measure along with the experiments and analysis is presented in Section 5. Significance of link prediction and the related review of the methods are also provided. Finally we conclude with summary of findings in Section 6.

2 Problem Statement

The social relation between two nodes has been viewed here as a string connecting the nodes (actors). The problems dealt with are as follows:

- To identify and quantify the tension force and its elementary component forces acting on a social relation by exploiting only the structural information of the network
- To design a link prediction algorithm as an application of the tension measure

3 Double Bounded Rough Set

Before describing in Section 4 the details of the forces acting between actors, here we will define the mathematical foundations that would be used in developing the underlying theory. In modern social networks, relations show complex characteristics due to high overlap in the neighborhood. Hence, it is very hard to define crisp boundary of a node's neighborhood. Expressing such ill-defined neighborhood in granular computing framework provides advantages in problem solving [6], [7]. A granule is a collection of data points which are indiscernible with respect to a given set of attributes. In such an environment one may need to extract the knowledge about a relational tuple. A relational tuple, say (a, b), is a relation which indicates that a data point a is explicitly related to another point b. This relationship might be of different types, e.g., transaction in financial data, friendship in social network data, email exchange in communication data, co-participation in event related data. A granule is said to be related to a data point a when any element of the granule is explicitly related to a. One may define the domain of a tuple by all the granules which are related to the tuple. Granules in the domain may not be expressed with a crisp set as there might be granules which are related to only one of the member of the tuple instead of both the members. In other words, granules may partially fall within the domain of the said relationship. The theory of rough sets [9] appears to be appropriate to represent such imperfect knowledge. In rough set theoretic framework, the lower approximate region of the set consists of the granules which are related to both the players a and b of the tuple, and the upper approximate region contains the granules which are related to either of a and b, or both a and b. That means the region corresponding to "either of a and b" represents the boundary (possibly belonging) region of the set. One may note that, in case of relation data, "either of a" and "either of b" may characterize two different aspects of knowledge. Therefore, it may be noted that the boundary region, i.e., (upper - lower) approximate region, here, comprises two distinct classes of granules, namely, those related to only a and those related to only b. In order to deal with this situation, we propose here a new variant of rough set, namely, double bounded rough set, to express the domain of such relational tuple in the granulated environment.

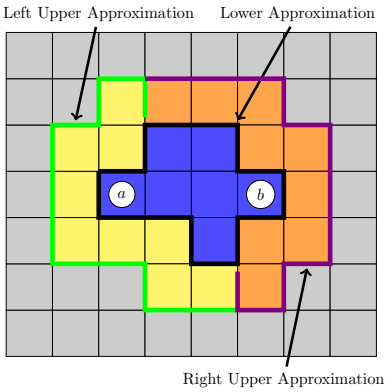


Fig. 1: Illustration: Double Bounded Rough Set

The double bounded rough set, defined here, has three distinct elements in the set. One is the lower approximation, second is the left upper approximation and the last one is the right upper approximation. Figure 1 shows an illustration of double bounded rough set. Here, a and b represents the two nodes between which an explicit relation exists. Blue region indicates the lower approximation of this relation. Blue and yellow together constitute the left-upper approximation and blue plus orange define the right-upper approximation of the relation. Let us now mathematically define the double bounded rough set.

Suppose we have an information system $S = (U, A)$, where U is the *universe* and A is the set of *attributes*. Both U and A are finite and non empty sets. For any $P \subseteq A$, there is an equivalence relation $IND(P)$ such that

$$IND(P) = \left\{ (x, y) \in U^2 \;\middle|\; \forall p \in P, p(x) = p(y) \right\} \tag{1}$$

Here, $p(x)$ function returns the value of the attribute p for data point x. The relation $IND(P)$ is called P-indiscernibility relation and any two points $(x, y) \in P$ indicate that x and y can not be distinguishable using the attribute set P. The equivalence class of the P-indiscernibility relation is denoted by $[x]_P$, and U/P denotes all the classes. Let us denote this granulated information system with $S_P = (U, A, [x]_P)$.

When dealing with the relational data, data points usually have explicitly defined connections (e.g. friendship, follower, transactions etc.) with each other. Let this connection be denoted by Q. In S_P, let us now define a relation I on $U \times U/P$ such that,

$$I(x) = \left\{ [x]_P \;\middle|\; \exists y \in [x]_P \text{ where } (x, y) \in Q \right\} \tag{2}$$

For a given information system $S_P = (U, A, [x]_P)$, and relations Q and I, we now define three operations assigning to every xQy. These three operations output three sets $P_*(xQy)$, ${}^*P(xQy)$ and $P^*(xQy)$, called P-lower, P-left-upper and P-right-upper approximation of (xQy) pair respectively. These are defined as follows:

$$P_*(xQy) = I(x) \cap I(y) \tag{3}$$
$${}^*P(xQy) = I(x) \cup (I(x) \cap I(y)) \tag{4}$$
$$P^*(xQy) = I(y) \cup (I(x) \cap I(y)) \tag{5}$$

where $I(x)$ returns the set of P-granules which are related to data point x.

Hence, P-lower approximation of the domain of relation xQy is the collection of P-granules which has I-relationship with both x and y data points. On the other hand, P-granules in the boundary region have I-relationship with either x or y. The boundary region of the set is defined by

$$({}^*P(xQy) \cup P^*(xQy)) \setminus P_*(xQy) \tag{6}$$

If the boundary region is empty then the set is crisp with respect to P, otherwise the set is rough. In addition, the upper approximation of the set has two different bounds. Hence, the set is named as double bounded rough set.

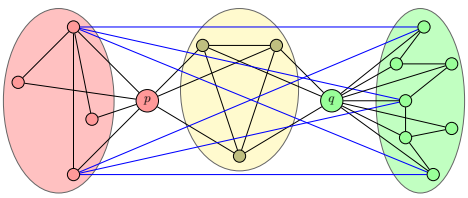


Fig. 2: An example network

4 Tension Between a Pair of Social Actors

In this section, we define a social tension measure of a pair of nodes in the network. The tension measure attributed to a pair could predict the dynamics between the two individuals in the pair.

A person's current network structure can reveal the current psychological state of that person with its neighboring peers. From these structures we could identify several forces which are critical to the changes in the network. The source of these forces is explained here before formulating the tension measure mathematically.

The first source of attraction (f_1) for a pair of nodes (say, (p, q)) is the common neighbors. In the example network (Figure 2), this region is shown with light yellow. Classical say is that larger the number of common neighbors, higher the probability that p and q are attracted towards each other. Our intuition says that a positive force towards the formation (or beholding) of connection between p and q comes from the activities within the common neighbors. Hence more denser the common neighbors are, more the nodes p and q are attracted to each other.

Second force (f_2), we identified, is due to the presence of inter-neighborhood communication links. This force is also a positive one towards the relational pair (p, q). Inter-neighborhood links are the links between the neighbors of only p with the neighbors of only q. For example in Figure 2, blue edges connects a node from light red region of p's neighbor to light green region of q's neighbor, i.e., friends of p is in friend with the friends of q. This puts a peer pressure in the relation (p, q) to come closer.

Not all the forces are positive for a relationship. There are activities which may push two nodes apart. One such force comes from the intra-neighborhood connectivity. It is possible that the neighborhood of a node (say, p) is densely connected with each other and this kind of higher density of intra-neighborhood connections will form a strong group within the node p's neighborhood. This factor repulses p to form a new connection with some other node q outside the group, as p is busy and satisfied with his/her existing friends. Therefore, such forces (f_3 and f_4) around both p and q contribute negatively towards their relationship. The regions of f_3 and f_4 are shown in Figure 2 with light red and light green overlays respectively.

Thus there are three different components of the neighborhood that generate four different forces for a relationship. Two of them are positive, while the other two are negative. A cumulative resultant tension force can quantify the strength of a relationship in a social network. In the next part, we mathematically formulate this tension force.

4.1 Network Model

The network is represented with a graph $G(V, E)$, where V is the set of nodes and E is the set of edges. The network is further granularized where each granule is a pair of nodes, say $\{x, y\}$ where $x, y \in V$. Let the set of all the granules in the network be denoted by $\mathcal{G}$. Consider a function $e(.)$ defined over a granule such that $e(\{x, y\}) = 1$ if $(x, y) \in E$, and zero, otherwise. The neighborhood of a node p is defined as $\Gamma(p) = \{\ g \in \mathcal{G} \mid p \in g, e(g) = 1\ \}$.

4.2 Different forces between two actors

As stated earlier, to analyze a local relationship, we need to see the neighborhood of the relationship. Here we express the neighborhood ($\mathcal{N}$) of a node pair (p, q) in the notion of the aforesaid double bounded rough set (Section 3) as

$$\mathcal{N}_*(p,q) = \{\ g \mid \forall x \in g, \{x,p\} \in \Gamma(p)\ \&\ \{x,q\} \in \Gamma(q)\ \} \quad (7)$$

$$\mathcal{N}^*(p,q) = \{\ g \mid \forall x \in g, \{x,q\} \in \Gamma(q)\ \} \quad (8)$$

$$^*\mathcal{N}(p,q) = \{\ g \mid \forall x \in g, \{x,p\} \in \Gamma(p)\ \} \quad (9)$$

Density of links in lower approximate region is the first force and we can quantify it as

$$f_1(p,q) = \frac{\left|\left\{\ g \in \mathcal{N}_*(p,q) \mid e(g) = 1\ \right\}\right|}{|\mathcal{N}_*(p,q)|} \quad (10)$$

The numerator counts the number of granules in the lower approximate region where a physical link in the network exists and the denominator is the count of all the possible granules in the lower approximate region of the pair p and q.

The second force (f_2) is defined by the density of links within the inter-neighborhood connections. In computing this, we avoid the links within the common neighbors as they are already considered by f_1. Hence, f_2 is defined as in Equation 11.

$$f_2(p,q) = \frac{\left|\left\{\ g \mid x \in \bigcup_{p \in {}^*\mathcal{N}(p,q) \setminus \mathcal{N}_*(p,q)} p\ \wedge\ y \in \bigcup_{q \in \mathcal{N}^*(p,q) \setminus \mathcal{N}_*(p,q)} q \textbf{ where } g = \{x,y\};\ e(g) = 1\ \right\}\right|}{\left|\left\{\ g \mid x \in \bigcup_{p \in {}^*\mathcal{N}(p,q) \setminus \mathcal{N}_*(p,q)} p\ \wedge\ y \in \bigcup_{q \in \mathcal{N}^*(p,q) \setminus \mathcal{N}_*(p,q)} q \textbf{ where } g = \{x,y\}\ \right\}\right|} \quad (11)$$

The denominator of f_2 counts the granules for which one element is in the left boundary region and the other element is in the right boundary region, and the numerator counts how many granules of the denominator corresponds to a physical link in the network.

The other forces, f_3 and f_4 are due to the link density within the neighborhood of the participating pairs. These intra-neighborhood forces are quantified by the ratio of the number of granules in the left (or right) boundary region having physical link in the network, and the total number of granules in the left (or right) boundary region. The mathematical formulae of f_3 and f_4 is as follows.

$$f_3(p,q) = \frac{\left|\left\{\ g \in {}^*\mathcal{N}(p,q) \setminus \mathcal{N}_*(p,q) \mid e(g) = 1\ \right\}\right|}{\left|\left\{\ g \mid g \in {}^*\mathcal{N}(p,q) \setminus \mathcal{N}_*(p,q)\ \right\}\right|} \quad (12)$$

$$f_4(p,q) = \frac{\left|\left\{\ g \in \mathcal{N}^*(p,q) \setminus \mathcal{N}_*(p,q) \mid e(g) = 1\ \right\}\right|}{\left|\left\{\ g \mid g \in \mathcal{N}^*(p,q) \setminus \mathcal{N}_*(p,q)\ \right\}\right|} \quad (13)$$

4.3 Tension Measure ($\mathcal{T}$)

As real world social networks are psychologically driven, a deep psychology works behind the creation of new relations and breaking up of old relations in them. The density terms, discussed before also incorporate the relational psychology. The density component f_1 and f_2 contribute positively towards strong relationship, but higher density value of f_3 and f_4 works negatively for the relation. The combination of the four density terms results in the relational tension strength of the pair of nodes p and q.

Tension $\mathcal{T}$ between a social pair (p, q) is defined as

$$\mathcal{T}(p,q) = \frac{1}{2} \times [f_1(p,q) + f_2(p,q) - f_3(p,q) - f_4(p,q)] \quad (14)$$

Algorithm 1 shows a method to calculate the tension value.

4.3.1 *Characteristics*

- The value of $\mathcal{T}(p, q)$ varies from -1 to $+1$.
- $\mathcal{T}(p, q) = 1$ when $f_1 = 1$ and $f_2 = 1$, but f_3 and f_4 are zero (0). Physically it means that all the common neighbors are linked with each other and inter-neighbors are fully connected, but there is no intra-neighborhood connections. Figure 3 shows an example network with the said situation.

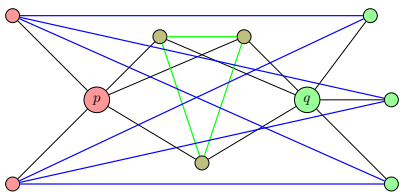


Fig. 3: Network with $\mathcal{T}(p, q) = 1$

- When $f_1, f_2 = 0$ and $f_3, f_4 = 1$, the value of $\mathcal{T}(p, q)$ attains the value -1. In other words, the tension is -1 when there is no connections within the common

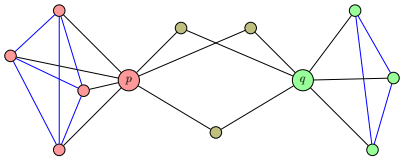


Fig. 4: Network with $\mathcal{T}(p,q) = (-1)$

neighbors and between inter-neighbors, but intra-neighborhoods of both p and q are fully connected. Figure 4 illustrates a situation where $\mathcal{T} = -1$.

- Tension measure along the pair (p,q) would be zero under three different conditions:

  1) When all the density components are zero (0), i.e., $f_1, f_2, f_3, f_4 = 0$, practically, it means the nodes are way apart from each other.
  2) When $f_1 + f_2 = f_3 + f_4$, indicating a neutral strength of the relation between the pair, and
  3) When all the forces attain their highest value of unity, i.e., $f_1, f_2, f_3, f_4 = 1$. That is the relation is neutral and saturated.

4.3.2 Observations in Synthetically Generated Networks

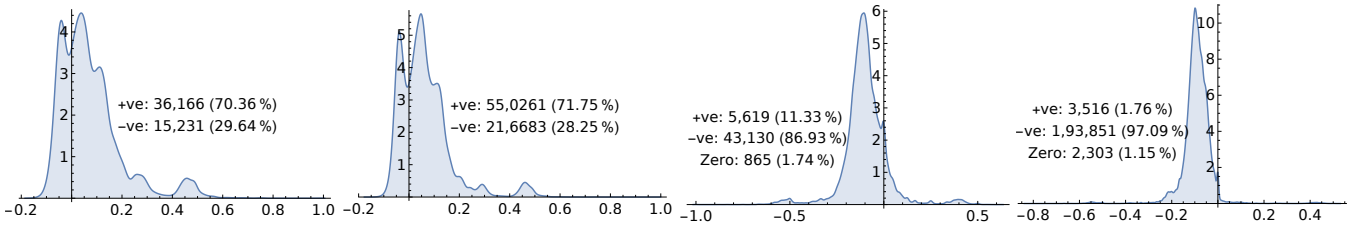
We generated two synthetic networks using LDBC [10]. The details of these data sets are shown in Table 1. In each network we calculated the tension measure for all the pairs with physical edges in the network. Figures 5a and 5b show the distributions of the value of $\mathcal{T}$ for the Data Set 1 and Data Set 2 respectively. We found that for Data Set 1, out of $51,398$ links $36,166$ links have $\mathcal{T}$ value greater than 0, i.e., the network has 70.36% of links which have positive tension. For Data Set 2, it is 71.75% positive and 28.25% negative values for $\mathcal{T}$. A tension measure shows the strength in a relationship. Hence, it is expected that those having physical link in the network will have a positive $\mathcal{T}$. This is also verified, to be true as we obtained higher percentage of positive values for existing links experimentally. However, there are negative values as well. For a dynamic social network, links may change their status quo over time. Negative value of tension for any physical link of such network indicates that the link became inactive and it may be removed from the network in future.

The distributions of four different forces are shown in the Figures 6a and 6b. For both the networks most of the f_1 values are distributed within the range of 0 to 0.5, whereas the values of other three forces f_2, f_3 and f_4 are distributed between 0 to 0.2.

We also experimented with the non-linked pairs. For these we choose random pairs of nodes of size $49,614$ and $1,99,670$ respectively for Data Set 1 and Data Set 2. The distributions of tension measure of these non-link pairs are shown in Figures 5c and 5d. As expected most of these values are negative; to be precise, it is 86.93% and 97.09% negative for Data Set 1 and Data Set 2 respectively. Even though they are not physically connected, about 11% and 2% (Figures 6c and 6d) nodes show positive tension value depicting a possibility of link formation in future.

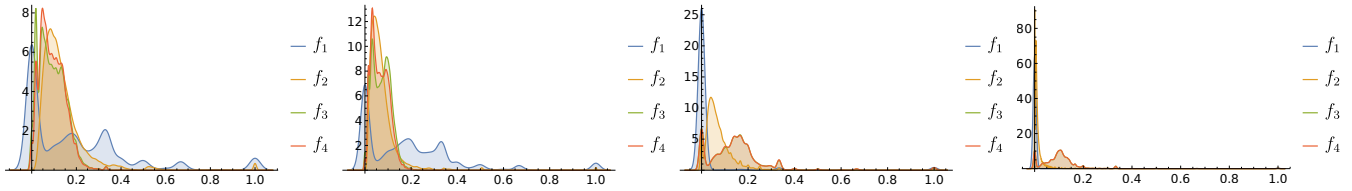
Algorithm 1 Tension calculation

```
'''G: networkx graph; a, b: nodes'''
def GetTensionValue(G, a, b):
  ten = 0.0, c_nbrs = set([])
  l_nbrs = [], r_nbrs = []

  '''identifying nodes in the
  rough regions of the neighborhood'''
  tmp = set(G.neighbors(b))
  for n in G[a]:
    if n in tmp:
      c_nbrs.add(n)
    else:
      l_nbrs.append(n)
  r_nbrs = tmp - c_nbrs

  '''Inter boundary density'''
  f2 = 0, count = 0
  for n in l_nbrs:
    x = set(G[n]).intersection(r_nbrs)
    count = count + len(x)

  ncl = len(l_nbrs) * len(r_nbrs)
  if (ncl > 0):
    f2 = float(count)/ncl

  '''common neighborhood density'''
  count = 0, f1 = 0
  if (len(c_nbrs) > 1):
    cmn = set(c_nbrs)
    while cmn:
      node = cmn.pop()
      x = set(G[node]).intersection(cmn)
      count = count + len(x)

    f1 = float(count)/ncr(len(c_nbrs), 2)

  '''Intra neighbor density @ src'''
  count = 0, f3 = 0
  if (len(l_nbrs) > 1):
    srcn = set(l_nbrs)
    while srcn:
      node = srcn.pop()
      x = set(G[node]).intersection(srcn)
      count = count + len(x)

    f3 = float(count)/ncr(len(l_nbrs), 2)

  '''Intra Neighbor Density @ dst end'''
  count = 0, f4 = 0
  if(len(r_nbrs)>1):
    dstn = set(r_nbrs)
    while dstn:
      node = dstn.pop()
      x = set(G[node]).intersection(dstn)
      count = count + len(x)
    f4 = float(count)/ncr(len(r_nbrs), 2)

  ten = 0.5*(f1 + f2 - f3 - f4)
  return ten
```



(a) Linked pairs of Data Set 1 (b) Linked pairs of Data Set 2 (c) Non-link pairs of Data Set 1 (d) Non-link pairs of Data Set 2

Fig. 5: Distribution of $\mathcal{T}$



(a) Linked pairs in Data Set 1 (b) Linked pairs in Data Set 2 (c) Non-linked pairs in Data Set 1 (d) Non-linked pairs in Data Set 2

Fig. 6: Distribution of different forces

4.3.3 *Observation in Real World Networks*

We have observed the behavior of the tension measure for pairs of nodes of five real world data sets including two network data collected by us, namely *Flickr* friendship network and *Twitter* Mention/Reply network. Features of these data sets are shown in Table 2. In the experiment we examined all the edged pairs for *Wiki-Vote* [11], [12] and *Facebook* [13] networks, whereas it is 50000 randomly selected edged pairs for *Flickr*, *Twitter* and *Youtube* [14] data. Obtained results are shown in Table 3.

TABLE 1: Synthetic Data Sets

	Data Set 1	**Data Set 2**
Nodes	3708	31444
Edges	51398	766948
Avg. Degree	27.7228	48.7818

It is evident from Table 3 that above 91% of the linked pairs have positive tension value indicating their relational tension strength positive. These results are as expected, and it justifies the significance of the tension measure. Nearly $6-9\%$ of the linked pairs were found to have tension values less than zero. These linked pairs, that reside in negative side of the distribution, indicate the possibility of their removal or deletion in future. The distribution plots of the linked pairs for all the data sets are shown in Figure 7.

We have also investigated the tension measure for non-existing edged pairs. A set of $50,000$ non-linked node pairs were selected randomly from all the data sets and we measured the tension score for all these links. The distributions of tension measure for different data sets are plotted in Figure 8. As expected, most of the tension values are either less than or equal to zero. The data summarized in Table 3 show positive tension values for less than 4% in case of facebook, youtube and twitter data, and less than 14% in case of wiki-vote and flickr data. These positive tension values indicate a possible future addition in physical links.

Thus, experimentally we found that for linked pairs, the tension values are mostly positive, whereas for non-linked pairs the tension is mostly either negative or zero. These findings validate the theoretical claims (4.3) of the tension measure. However, a major question arises from the aforesaid experimental observation that 'why there is a high percentage (over 40%) of zero in tension values for 3 out of 5 data sets?'. Even for wiki-vote, more than 15% pairs have tension measure zero. To understand the root cause of this results one may dig into the properties of the network. It is found that all these data sets contain high number (32.53% for wiki-vote, 48.63% for twitter, 59.69% for flickr and 52.88% for youtube) of nodes with degree less than 2. So, if a node of the pair has degree below two, then there would be less possibility of inter-connection in neighborhood. Hence, the structural tension value would be zero or indeterminable. We have also observed from the global aspect that, this situation may arise if the clustering coefficient of the whole network is very low. For example, here the average clustering coefficient (CC) of the Facebook network is 0.58, whereas for wiki-Vote, flickr, youtube and twitter the CC values are $0.14, 0.176, 0.082$ and 0.155 respectively.

5 Link Prediction

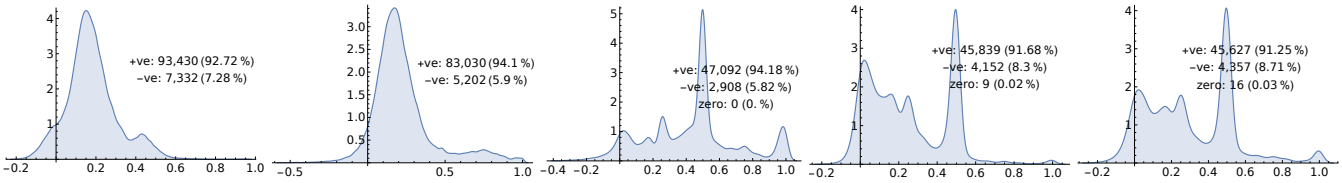
The problem of *link prediction* is to predict future links between a pair of social actors for a given snapshot of the network. It is an important problem to study as identifying such missing links may help in solving many important questions such as whether a person would like a book, a place or a picture or not, whether two scholars will collaborate with each other or not, and whether a person would receive some endorsement or not. Besides these academic interests, link prediction has many commercial applications as well, e.g., recommending friends in social network like facebook or finding probable job applicant in LinkedIn. It can also help in identifying the hidden groups in terrorist network or inferring missing links in taxonomies [15].

TABLE 2: Real World Data Sets

	Wiki-Vote	Facebook	Flickr	Youtube	Twitter	Haggle	Digg	Epinion	EUEmail
Nodes	7115	4039	559136	1076748	80521	274	279630	131828	986
Edges	103689	88234	1991356	2826692	219285	2124	1548126	711783	16064
Avg. Degree	28.3238	43.6910	7.1230	5.2504	5.4467	15.5036	11.0727	10.7987	32.5842
Nodes in Largest CC	7066	4039	559134	1073152	73126	274	261489	119130	986
Edges in Largest CC	103663	88234	1991355	2823873	213817	2124	1536577	704572	16064
Avg. Clustering Coefcient	0.1409	0.6055	0.17605	0.08276	0.15502	0.63268	0.09244	0.12794	0.40705
Temporal Info	No	No	No	No	No	Yes	Yes	Yes	Yes

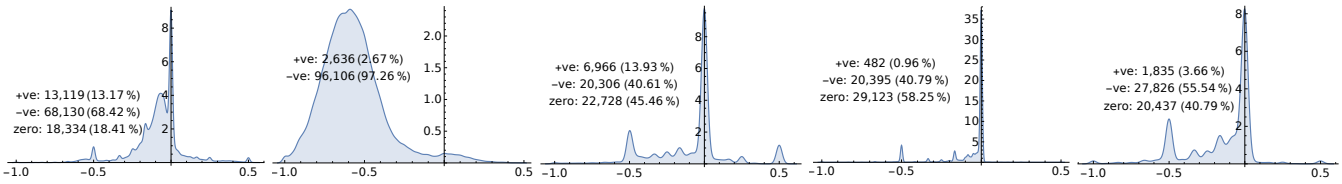
TABLE 3: Tension for linked and unlinked pairs

	Edged Pair			Non-Edged Pair		
Data Set	**+ve% (count)**	**-ve% (count)**	**0% (count)**	**+ve% (count)**	**-ve% (count)**	**0% (count)**
Wiki-Vote	92.72 (93430)	7.28 (7332)	0.00 (0)	13.17 (13119)	68.42 (68130)	18.41 (18334)
Facebook	94.10 (83030)	5.90 (5202)	0.00 (0)	2.67 (2636)	97.26 (96106)	0.07 (69)
Flickr	94.18 (47092)	5.82 (2908)	0.00 (0)	13.93 (6966)	40.61 (20306)	45.46 (22728)
Youtube	91.68 (45839)	8.30 (4152)	0.02 (9)	0.96 (482)	40.79 (20395)	58.25 (29123)
Twitter	91.25 (45627)	8.71 (4357)	0.03 (16)	3.66 (1835)	55.54 (27826)	40.79 (20437)



(a) Wiki-Vote (b) Facebook (c) Flickr (d) Youtube (e) Twitter

Fig. 7: Distribution of $\mathcal{T}(p, q)$ where $e(p, q) = 1$



(a) Wiki-Vote (b) Facebook (c) Flickr (d) Youtube (e) Twitter

Fig. 8: Distribution of $\mathcal{T}(p, q)$ where $e(p, q) = 0$

5.1 Related Work

Having multi-disciplinary applications, the problem of link prediction has been in active research for couple of decades. Generally the principle of link prediction algorithms is to assign a likelihood score to each of the non-existing links, order them according to their likelihood score [16] and select some of the top ones as per the need. The assumption here is that the links with the higher score have higher probabilities to be a link in future [17]. Scholars used different local and global characteristics to calculate the likelihood values. Based on the topological properties used, one may categorize the available algorithms in three [18] different sections viz, those using (i) neighbor based metric (ii) path based metric and (iii) random walk based metric. It is considered that people create links when they are close, and a well known assumption is that two unknown persons may become known due to their neighbors. Hence, people have designed many similarity metrics based on the neighborhood of the nodes. Examples of such measures are Common Neighbors [19], Preferential Attachment [20], Hub Promoted [21], Adamic-Adar [22], Jaccard Coefficient [23], Leicht-Holme-Nerman [24], Hub Depressed [25] and Parameter-Dependent [26]. There are methods where the path between two nodes has been used instead of their neighborhoods for measuring the similarities between them. These are the path based likelihood metrics. Katz measure [27], Local Path Score [28], Relation Strength Similarity [29] and FriendLink [30] are some of the popular path based link prediction metrics. Another way to model social interactions is the random walk algorithms where transition probabilities are used to determine the destination of the random walk from the current node. Based on these algorithms some of the link prediction metrics like SimRank [31], Hitting Time [32], PropFlow [33] and Rooted PageRank [33] have been developed.

More recently, researchers are trying to incorporate other social network theories along with the topological information in solving the link prediction problem. Some of the techniques developed in this approach are based on combining community level feature [34], developing individual relational network [35], and combining non-topological information such as user profile and geographical locations

[36].

Many learning based link prediction methods using the similarity matrices, non-topological attributes and external information have also been reported in recent years. The learning framework as used was supervised [37], [38], [39], [40], semi-supervised [40] and unsupervised [41], [42].

5.2 Algorithm based on the proposed tension measure

In this section we propose a link prediction algorithm using the tension measure defined in Section 4.3. Unlike the existing popular link prediction algorithms, where a similarity score is assigned to a node pair, here the tension measure assigns a direction (positive or negative) along with the similarity score. If the tension measure is positive then we predict a link formation, whereas if the score is negative then a link deletion is predicted. In other words, our link prediction algorithm calculates the tension measure for all the test pairs of nodes and then select the pairs with positive tension value. Unlike the other algorithms, we do not rank the predicted links because any positive tension value (irrespective of its magnitude) refers as a possibility of link formation between the concerned pair. The method is listed in Algorithm 2.

Algorithm 2 Link prediction using Tension measure

```
'''input for the program is a
graph and list of test pairs'''
def GetPredictedLinks(G, test_pairs):
  output = {}

  for (x,y) in test_pairs:
    t = GetTensionValue (G,x,y)
    if t > 0:
      output[(x,y)] = t

  return output
```

5.3 Descriptions of Data Sets

In our experiments we used two types of real world network data (i) without temporal information and (ii) with temporal information.

- **Wiki-Vote:** The data set was collected by SNAP group and publicly available in their homepage [43]. The network contains all the Wikipedia voting data till January 2008. Nodes represent Wikipedia users and the connections represent who vote for whom. The network is a directed one. However, we removed here the direction and used the network as a undirected network.
- **Facebook:** The source of the data is again SNAP [43] group and the network contains the ego net of Facebook users. It is an undirected network of 4039 users.
- **Youtube:** The youtube data was collected by OSNR group of The Max Planck Institute of Software Systems and publicly available at [44]. The properties are listed in Table 2.
- **Flickr:** The data was collected by us in our laboratory. The data contains the Flickr user-to-user links crawled during November and December 2015. The crawl was initiated by seeds of Indian origin in Flickr. Total link crawled was about 2 million and the features of the network are shown in Table 2.
- **Twitter:** The network is formed by the reply/mentions of tweets collected by us during January to July 2017. The network nodes represent the twitter users and a link is added if an user is either replied or (s)he mentions another name in her (his) tweet. The features are listed in Table 2.
- **Haggle [45], [46]:** This data set is an undirected temporal network of contacts between people within wireless devices. A node represents a person and an edge between two persons shows that there was a contact between them.
- **Digg [47], [48]:** The temporal data set is a friendship network of digg users collected in 2009.
- **Epinion [49], [50]:** This is a temporal trust network of epinion, an online product rating site. The network consist of individual users connected by directed trust links.
- **EUEmail:** The network was generated using email data from a large European research institution by SNAP [43] group. An edge of this network represent a email communication between the members.

5.4 Comparing Methods

We compared our proposed link prediction method based on tension measure with different popular similarity metrics based link prediction algorithms []. The objective is to demonstrate the effectiveness of the proposed tension measure vise-a-vise the existing similarity matrices. The different similarity metrics used so far are as follows:

- **Common Neighbors:** The common neighbor predictor assigns a similarity score to a pair of nodes based on the number of common neighbors between them. The philosophy behind the common neighbor predictor is that a common friend may introduce the two parties. Hence more the common friends are, higher is the chance of them to become friend. Common neighbor score $CN(x, y)$ of x and y is calculated as

$$CN(x, y) = \Gamma(x) \cap \Gamma(y) \tag{15}$$

  where $\Gamma(.)$ represents the neighbors of $(.)$.
- **Resource Allocation Index:** The resource allocation index gives different weights to different common neighbors. It says, lesser the degree of a common neighbor (p) of two nodes (a and b), higher the chance of that pair (a, b) to be friend through p. The resource allocation index is calculated by the sum of inverse of the degree of each of the common neighbors as

$$RA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|} \tag{16}$$

- **Adamic Adar:** Adamic Adar is a further modification of resource allocation index and is calculated by

$$AA(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|} \tag{17}$$

- **Preferential Attachment:** Based on the scale free properties of social networks, it is a well known fact that a user with many friends tend to create more future connections. Based on this, the preferential attachment score is calculated as

$$PA(x,y) = |\Gamma(x) \times \Gamma(y)| \tag{18}$$

- **Jaccard's Coefficient:** This is widely used in information retrieval. The measure is capable of comparing the similarity as well as the diversity in the neighborhood of the concerning node pair. The Jaccard's coefficient is measured by

$$JC(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{19}$$

5.5 Experimental Setup

We have conducted experiments to compare the performance of different similarity matrices (Equations (16) to (19)) with the proposed tension measure (Equation (14)). The experiments were executed over two different types of network data, namely, those with temporal information and those without temporal information.

For the data which do not have temporal information, we have selected 10% of the links randomly and kept them as the prob set (say, E'_{prob}), and the remaining 90% links were used for training. The graph generated from the training links was used as the input set of both proposed and comparing algorithms for link prediction. Apart from the prob set of 10% existing edges, we also randomly selected non-lined pair (say, E''_{prob}) of similar size. Thus the total link pair selected as prob set is $E_{prob} = E'_{prob} \cup E''_{prob}$. However, for the temporal data the prob set is selected based on the temporal information. With these prob sets, following quantitative and statistical indexes were computed for comparative study.

- **Effectiveness ($\mathscr{E}$):** One way to determine the effectiveness of a predictor is to measure the percentage of relevant objects correctly identified, i.e., the true positive (tp). However, in the process of maximizing tp, a predictor may select non-relevant objects, i.e., the false positive (fp). So, we measured the effectiveness, with the difference between the true positive rate and false positive rate as,

$$\mathscr{E} = \frac{tp}{tp+fn} - \frac{fp}{tn+fp} \tag{20}$$

where fn is the false negative and tn is the true negative.

- **Accuracy ($\mathscr{A}$):** Accuracy measure of a predictor is

$$\mathscr{A} = \frac{tp+tn}{tp+fp+tn+fn} \tag{21}$$

- **Area under the receiver operating Characteristics curve [51], [52]:** This measure can be interpreted as the probability that a randomly chosen missing link is given a higher score than a randomly chosen non-linked pair. Considering n independent comparisons, and the accuracy is measured as

$$AUC = \frac{n' + 0.5n''}{n} \tag{22}$$

where n' is the number of times a missing link is given higher score and n'' is the number of times a missing link is given equal score to that of the non-linked pair.

- **f-score:** The measure f-score is the harmonic mean of precision and recall. Precision is defined as the factor of identified links that were missing link. On the other hand, recall is the factor of the missing link selected. Hence, higher the value of f-score, better the predictor. f-score is measured as

$$f = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{23}$$

where precision and recall are

$$Precision = \frac{tp}{tp+fp} \tag{24}$$

$$Recall = \frac{tp}{tp+fn} \tag{25}$$

5.6 Results

Table 4 shows the comparative values of the measure Effectiveness $\mathscr{E}$ for different data sets. For all the data sets tension based link prediction algorithm shows highest effectiveness-value. For example, the improvement over the nearest algorithm is 12.95% (from JC score) in wiki-vote, 4.29% (from RA score) in facebook, 90.25% (from AA score) in flickr, 11.89% (from AA score) in youtube and 24.23% (from AA score) in twitter. This indicates that the proposed method is superior in selecting the missing links while rejecting the falls positive, with $\mathscr{E}$ at least 10% better than others, except the facebook data where tension is superior by 4.3%. A graphic showing the box whiskers plot of the obtained results is shown in Figure 9 for the measure $\mathscr{E}$.

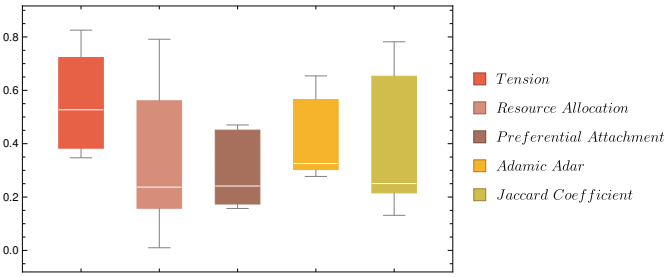


Fig. 9: Effectiveness of different data sets for different algorithms. The white bar within the box shows the median value and the upper and lower boundary of the box are the upper and lower quartiles of the obtained values. The whiskers show the highest and lowest values reported.

$\mathscr{A}$ values of different algorithms for different data sets are listed in Table 5. Similar to the $\mathscr{E}$-index, the proposed prediction method with $\mathcal{T}$-measure, shows improvement

TABLE 4: Comparison in terms of $\mathscr{C}$

	wiki-vote	Facebook	Flickr	Youtube	Twitter	Haggle	Digg	Epinion	EUEmail
Tension	0.689	0.826	0.527	0.347	0.395	0.330	0.235	0.294	0.546
Resource Allocation	0.484	0.792	0.010	0.207	0.237	0.014	0.127	0.199	0.003
Preferential Attachment	0.470	0.444	0.241	0.157	0.179	0.175	0.195	0.292	0.368
Adamic/Adar	0.536	0.654	0.277	0.311	0.326	0.108	0.198	0.178	0.459
Jaccard Coefficient	0.610	0.782	0.132	0.251	0.244	0.076	0.159	0.246	0.435

TABLE 5: Comparison in terms of $\mathscr{A}$

	wiki-vote	Facebook	Flickr	Youtube	Twitter	Haggle	Digg	Epinion	EUEmail
Tension	0.845	0.913	0.764	0.674	0.698	0.672	0.618	0.647	0.774
Resource Allocation	0.744	0.897	0.505	0.604	0.619	0.579	0.564	0.599	0.512
Preferential Attachment	0.737	0.724	0.621	0.579	0.590	0.648	0.598	0.646	0.689
Adamic/Adar	0.769	0.829	0.639	0.656	0.663	0.620	0.599	0.589	0.735
Jaccard Coefficient	0.806	0.892	0.566	0.625	0.622	0.604	0.580	0.623	0.723

TABLE 6: Comparison in terms of AUC

	wiki-vote	Facebook	Flickr	Youtube	Twitter	Haggle	Digg	Epinion	EUEmail
Tension	0.860	0.991	0.827	0.602	0.782	0.710	0.520	0.650	0.830
Resource Allocation	0.922	0.995	0.826	0.734	0.783	0.730	0.660	0.765	0.925
Preferential Attachment	0.937	0.833	0.840	0.763	0.800	0.660	0.685	0.880	0.750
Adamic/Adar	0.927	0.993	0.817	0.721	0.778	0.770	0.675	0.790	0.870
Jaccard Coefficient	0.918	0.989	0.804	0.725	0.769	0.660	0.610	0.800	0.850

TABLE 7: Comparison in terms of f-Score

	wiki-vote	Facebook	Flickr	Youtube	Twitter	Haggle	Digg	Epinion	EUEmail
Tension	0.838	0.907	0.735	0.522	0.588	0.616	0.413	0.476	0.760
Resource Allocation	0.656	0.884	0.020	0.343	0.384	0.028	0.225	0.332	0.006
Preferential Attachment	0.648	0.642	0.389	0.272	0.305	0.297	0.328	0.452	0.597
Adamic/Adar	0.703	0.792	0.434	0.475	0.492	0.196	0.332	0.302	0.648
Jaccard Coefficient	0.769	0.878	0.277	0.402	0.411	0.162	0.279	0.399	0.623

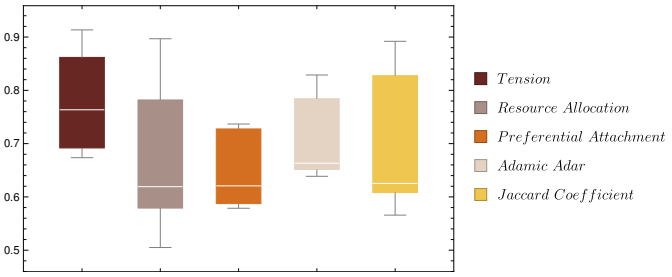


Fig. 10: Accuracy of different data sets for different algorithms. The white bar within the box shows the median value and the upper and lower boundary of the box are the upper and lower quartiles of the obtained values. The whiskers show the highest and lowest values reported.

over the best obtained values of 'accuracy' by other methods for all the data sets. The factor of improvement from the highest score obtained by the existing algorithm is 4.83%, 2.35%, 19.56%, 2.74% and 5.28% respectively for wiki-vote, facebook, flickr, youtube and twitter data. Figure 10 shows a plot of the obtained results.

AUC score indicates whether the obtained values for missing links are usually higher than the obtained values of non-linked pairs. Table 6 shows the AUC values of all the data sets for all the matrices. Unlike accuracy and effectiveness, the proposed tension based link prediction algorithm shows mixed results. For facebook, the proposed method shows comparable results, whereas for flickr and twitter it is comparable with others, except the Preferential Attachment metric. For wiki-vote and youtube, the score is lowest for the tension measure as compared to the other comparing similarity matrices. A plot showing the obtained results is in Figure 11.

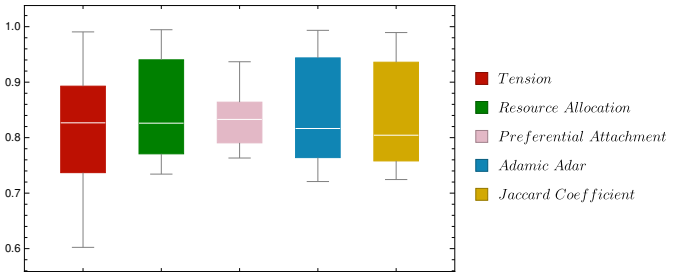


Fig. 11: AUC of different data sets for different algorithms. The white bar within the box shows the median value and the upper and lower boundary of the box are the upper and lower quartiles of the obtained values. The whiskers show the highest and lowest values reported.

Finally, the results of f-score are shown in Table 7. From the table it is clear that the f-score for the proposed method is high as compared to all the methods for all the data sets we have experimented with. It indicates that the proposed $\mathcal{T}$ measure can maximize both the precision and recall. That is, while being precise it can select more relevant items as compared to the other comparing matrices. Figure 12 summarizes the obtained f-scores in a box whiskers chart.

All the aforesaid experiments have been executed in an Intel Core i5-5200U at 2.20 GHz based system with 8 GB

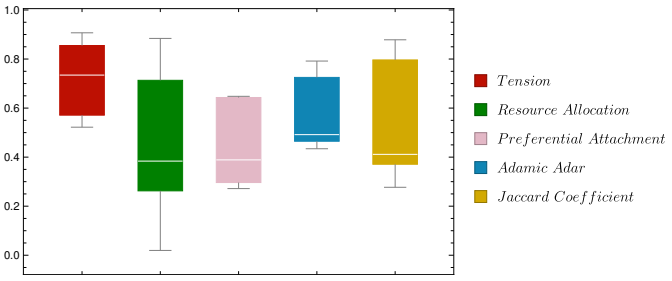


Fig. 12: f-score of different data sets for different algorithms. The white bar within the box shows the median value and the upper and lower boundary of the box are the upper and lower quartiles of the obtained values. The whiskers show the highest and lowest values reported.

DDR4 RAM. The programming language used is python 3.6 under fedora core 25. Multiprocessor coding has been used to reduce the execution time of all the experiments. The comparative execution times (for few data sets as example) are shown in Figure 13. Each group corresponds to a particular list of test pairs. 'nl' in the graph label refers to the non link (E''_{prob}) prob set. Execution time-wise, the method based on the proposed $\mathcal{T}$ is the slowest among the all. However, one should note that the information content of the proposed tension measure is more than those of the other comparing matrices. For example, the proposed method can easily predict the link removal along with its ability in predicting future links, which is not possible by others. Another observation from Figure 13 is that for all the data sets the execution time of the proposed method for non linked pairs is significantly lower than that of the missing link pairs. On the other hand, the comparing methods take similar execution time for missing and non lined pairs for most of the data sets.

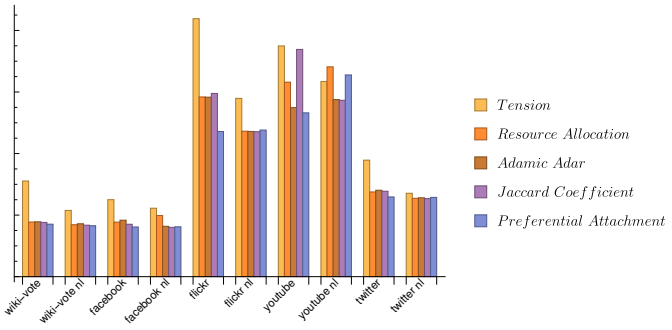


Fig. 13: Execution time in log

6 Discussion and Conclusion

The contribution of the current investigation is three fold. First, the social relation between two actors (nodes) is viewed as a string with various forces acting on it. Accordingly, a tension measure which pulls two participating nodes along the string is defined. Four different components of the tension measure are identified. A new varient of rough set, namely, double bounded rough set is introduced in order to quantify the said forces mathematically. Finally, an algorithm for link prediction in dynamic social network is developed based on the proposed tension measure.

Double bounded rough set (Section 3) concerns with the domain of a crisp relationship between two actors. It hinges on the concept of two upper bounds (unlike Pawlaks rough set [9]) along with the lower bound. Though its application is shown here for social network analysis, it can be used for other domains with relational data.

We identified that at least four forces are there between any pair of actors in a social network. Two of the forces are positively contributing to the relation, whereas the other two negatively contribute. Positive force means pulling of the participating nodes towards each other, while the negative one indicates their repulsion. Experimentally, we showed that more than 90% linked and less than 15% non-linked pairs of nodes have positive tension value. It is also found that for all the cases above 85% tension value is either negative or zero for non-linked pairs. It means the tension measure has a strong correlation with the physical links of the network. This indicates the suitability of the tension measure as an index for link prediction. It is shown statistically that the link prediction algorithm designed based on the tension measure outperforms several well known similarity based link prediction algorithms for most of the data sets considered here.

Although the tension based link prediction method is relatively slow, it has other merits and its application is not limited to link prediction only. Its characteristics of identifying the negative strength of relationship provide additional benefits in social network analysis. With this ability, one can identify the links which are possible candidates for future removal. The tension measure, being capable of predicting addition as well as deletion of links, can well fit in modeling network evolution. It can further be used to generate synthetic social network data with more dynamic characteristics. Influence maximization is another area where this measure may find application in rejecting the inactive links while identifying the influencing actors.

Acknowledgments

The authors would like to thank Mr. Arindam Kundu for fruitful discussion about the problems in early stage of the research. Authors also would like to thank Mr. Rupsayar Das for providing us with twitter data set used in experimental analysis. Professor Sankar K. Pal acknowledges DAE Raja Ramanna Fellowship and Sir J. C. Bose National Fellowship (grant-in-aid) of the Government of India. The project was partially supported by The National Science Centre, Poland the research project no. 2016/23/B/ST6/01735.

References

[1] G. Kossinets and D. J. Watts, "Empirical Analysis of an Evolving Social Network," *Science*, vol. 311, no. 5757, pp. 88–90, 2006.

[2] W. J. Burk, C. E. Steglich, and T. A. Snijders, "Beyond dyadic interdependence: Actor-oriented models for co-evolving social networks and individual behaviors," *International Journal of Behavioral Development*, vol. 31, no. 4, pp. 397–404, jul 2007.

[3] T. Tylenda, R. Angelova, and S. Bedathur, "Towards time-aware link prediction in evolving social networks," in *Proc. of the 3rd Workshop on Social Network Mining and Analysis - SNA-KDD'09*. Paris, France: ACM Press, 2009, pp. 9:1–9:10.

[4] A. D. Henry, P. Prałat, and C.-Q. Zhang, "Emergence of segregation in evolving social networks." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 21, pp. 8605–8610, may 2011.

[5] F. S. F. Pereira, "User Preferences Dynamics on Evolving Social Networks - Learning, Modeling and Prediction: Student Research Abstract," in *Proc. of the Symposium on Applied Computing*, ser. SAC '17. New York, USA: ACM, 2017, pp. 1090–1091.

[6] S. Kundu and S. K. Pal, "FGSN: Fuzzy Granular Social Networks - Model and applications," *Information Sciences*, vol. 314, pp. 100–117, 2015.

[7] ——, "Fuzzy-rough community in social networks," *Pattern Recognition Letters*, vol. 67, pp. 145–152, 2015.

[8] L. A. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets and Systems*, vol. 90, no. 2, pp. 111–127, 1997.

[9] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*. Norwell, MA, USA: Kluwer Academic Publishers, 1992.

[10] A. Prat, "DATAGEN: data generation for the Social Network Benchmark." [Online]. Available: http://ldbcouncil.org/blog/datagen-data-generation-social-network-benchmark

[11] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proc of the 19th international conference on World wide web - WWW '10*. New York, USA: ACM Press, 2010, pp. 641–650.

[12] ——, "Signed Networks in Social Media," in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. New York, New York, USA: ACM Press, 2010, pp. 1361–1370.

[13] J. J. McAuley and J. Leskovec, "Learning to discover social circles in ego networks," in *Proc. of 25th International Conference on Advances in Neural Information Processing Systems*, P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Lake Tahoe, Nevada, 2012, pp. 548–556.

[14] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks," in *Proc. of the 7th ACM SIGCOMM conference on Internet measurement*, San Diego, oct 2007, pp. 29–42.

[15] J. Liang, Y. Xiao, H. Wang, Y. Zhang, and W. Wang, "Probase+: Inferring missing links in conceptual taxonomies," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 6, pp. 1281–1295, jun 2017.

[16] H. Wang, W. Hu, Z. Qiu, and B. Du, "Nodes' evolution diversity and link prediction in social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 10, pp. 2263–2274, oct 2017.

[17] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[18] P. Wang, B. Xu, Y. Wu, and X. Zhou, "Link Prediction in Social Networks: the State-of-the-Art," vol. 58, no. January, pp. 1–38, 2014.

[19] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Physical Review Letters E.*, vol. 64, no. 2, pp. 251 021–251 024, 2001.

[20] A.-L. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaboration," *Physica A*, vol. 311, pp. 590–614, 2002.

[21] E. Ravasz, "Hierarchical Organization of Modularity in Metabolic Networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[22] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, no. 3, pp. 211–230, 2003.

[23] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, feb 1912.

[24] E. A. Leicht, P. Holme, and M. E. Newman, "Vertex similarity in networks," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 73, no. 2, 2006.

[25] T. Zhou, L. Linyuan, T. Zhou, and L. Linyuan, "Predicting missing links via local information Predicting missing links via local information," vol. 630, pp. 623–630, 2009.

[26] Y. X. Zhu, L. Lü, Q. M. Zhang, and T. Zhou, "Uncovering missing links with cold ends," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 22, pp. 5769–5778, 2012.

[27] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[28] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Physical Review E*, vol. 80, no. 4, p. 046122, 2009.

[29] H.-H. Chen, L. Gou, X. Zhang, and C. Giles, "Discovering missing links in networks using vertex similarity measures," in *Proc. of the 27th Annual ACM Symposium on Applied Computing (SAC'12)*, Trento, Italy, 2012, pp. 138–143.

[30] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos, "Fast and accurate link prediction in social networking systems," *The Journal of Systems & Software*, vol. 85, no. 9, pp. 2119–2132, 2012.

[31] G. Jeh and J. Widom, "SimRank : A Measure of Structural-Context Similarity," in *Proc. of the eighth ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD'02)*, Edmonton, Canada, 2002, pp. 538–543.

[32] F. Fouss, A. Pirotte, J. M. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 355–369, 2007.

[33] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proc. of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD'10*, Washington DC, USA, 2010, p. 243.

[34] A. De, S. Bhattacharya, S. Sarkar, N. Ganguly, and S. Chakrabarti, "Discriminative Link Prediction using Local, Community, and Global Signals," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2057–2070, aug 2016.

[35] S. Yan, K. J. Lin, X. Zheng, W. Zhang, and X. Feng, "An approach for building efficient and accurate social recommender systems using individual relationship networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 10, pp. 2086–2099, oct 2017.

[36] Z. Wang, J. Liang, R. Li, and Y. Qian, "An Approach to Cold-Start Link Prediction: Establishing Connections between Non-Topological and Topological Information," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 11, pp. 2857–2870, nov 2016.

[37] M. Pujari and R. Kanawati, "A Supervised Machine Learning Link Prediction Approach for Tag Recommendation," *Online*, pp. 336–344, 2011.

[38] H. Kashima and N. Abe, "A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction," in *Proc. of 6th International Conference on Data Mining, ICDM'06*, Washington DC, USA, 2006, pp. 340–349.

[39] M. Pujari and R. Kanawati, "Link prediction in complex networks by supervised rank aggregation," in *Proc. of 24th International Conference on Tools with Artificial Intelligence ICTAI*, vol. 1, Athens, Greece, 2012, pp. 782–789.

[40] E. Tasnádi and G. Berend, "Supervised Prediction of Social Network Links Using Implicit Sources of Information," in *Proc. of the 24th International Conference on World Wide Web - WWW '15 Companion*, Florence, Italy, 2015, pp. 1117–1122.

[41] T.-T. Kuo, R. Yan, Y.-Y. Huang, P.-H. Kung, and S.-D. Lin, "Unsupervised link prediction using aggregative statistics on heterogeneous social networks," in *Proc. of 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD'13*, Chicago, USA, 2013, pp. 775–783.

[42] M. Kaya, M. Jawed, E. Bütün, and R. Alhajj, "Unsupervised Link Prediction Based on Time Frames in WeightedDirected Citation Networks," in *Trends in Social Network Analysis: Lecture Notes in Social Networks*, R. Missaoui, T. Abdessalem, and M. Latapy, Eds. Springer, Cham, 2017.

[43] J. Leskovec, "SNAP Large Network Data." [Online]. Available: https://snap.stanford.edu/data/index.html

[44] "Online Social Networks Research at The Max Planck Institute of Software Systems." [Online]. Available: http://socialnetworks.mpi-sws.org/data-imc2007.html

[45] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of Human Mobility on Opportunistic Forwarding Algorithms," *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, pp. 606–620, 2007.

[46] "Haggle network dataset – {KONECT}," apr 2017. [Online]. Available: http://konect.uni-koblenz.de/networks/contact

[47] T. Hogg and K. Lerman, "Social Dynamics of {Digg}," *EPJ Data Science*, vol. 1, no. 5, 2012.

[48] "Digg friends network dataset – {KONECT}," apr 2017. [Online]. Available: http://konect.uni-koblenz.de/networks/digg-friends

[49] P. Massa and P. Avesani, "Controversial Users Demand Local Trust Metrics: an Experimental Study on {epinions.com} Community," in *Proc. American Association for Artificial Intelligence Conf.*, 2005, pp. 121–126.
[50] "Epinions trust network dataset – {KONECT}," apr 2017. [Online]. Available: http://konect.uni-koblenz.de/networks/epinions
[51] A. Hanley and J. McNeil, "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, vol. 143, pp. 29–36, 1982.
[52] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, jun 2006.

Suman Kundu received B.Tech in Information Technology from West Bengal University of Technology, Kolkata, India in 2005, and M.E. degree in Software Engineering and Ph.D. degree in Engineering from Jadavpur University, Kolkata, India in the year 2009 and 2017 respectively. He was a senior research fellow at Center for Soft Computing Research, Indian Statistical Institute, Kolkata, India from 2010 to 2015. He also worked in software development company ZINFI Software Systems Pvt. Ltd., Kolkata, India from 2006 to 2010 and 2016 to 2018. Currently he is a Post-doctoral researcher at Wroclaw University of Science and Technology, Wroclaw, Poland. He published 4 journal and 2 conference articles in the area of social network analysis, granular computing, soft computing and distributed computing. Visit his website http://sumankundu.info for more information.

Sankar K. Pal (M'80−SM'84−F'93−LF'15) is a Distinguished Scientist and former Director of Indian Statistical Institute, and a former Chair Professor of Indian National Academy of Engineering. He is currently a Distinguished Professor Chair of Indian National Science Academy (INSA). He founded the Machine Intelligence Unit and the Center for Soft Computing Research: A National Facility in the Institute in Calcutta. He received a Ph.D. in Radio Physics and Electronics from the University of Calcutta in 1979, and another Ph.D. in Electrical Engineering along with DIC from Imperial College, University of London in 1982.

He worked at the University of California, Berkeley and the University of Maryland, College Park in 1986-87; the NASA Johnson Space Center, Houston, Texas in 1990-92 & 1994; and in US Naval Research Laboratory, Washington DC in 2004. Since 1997 he has been serving as a Distinguished Visitor of IEEE Computer Society (USA) for the Asia-Pacific Region, and held several visiting positions in Italy, Poland, Hong Kong and Australian universities.

Prof. Pal is a Life Fellow of the IEEE, and Fellow of the World Academy of Sciences (TWAS), International Association for Pattern recognition, International Association of Fuzzy Systems, International Rough Set Society, and all the four National Academies for Science/Engineering in India. He is a coauthor of twenty books and more than four hundred research publications in the areas of Pattern Recognition and Machine Learning, Image Processing, Data Mining and Web Intelligence, Soft Computing, Neural Nets, Genetic Algorithms, Fuzzy Sets, Rough Sets, Cognitive Machine and Bioinformatics. He initiated and promoted the soft computing research & teaching in India. He visited forty five countries as a Keynote/ Invited speaker or an academic visitor.

He has received the 1990 S.S. Bhatnagar Prize (which is the most coveted award for a scientist in India), 2013 Padma Shri (one of the highest civilian awards) by the President of India and many prestigious awards in India and abroad including the 1999 G.D. Birla Award, 1998 Om Bhasin Award, 1993 Jawaharlal Nehru Fellowship, 2000 Khwarizmi International Award from the President of Iran, 2000-2001 FICCI Award, 1993 Vikram Sarabhai Research Award, 1993 NASA Tech Brief Award (USA), 1994 IEEE Trans. Neural Networks Outstanding Paper Award, 1995 NASA Patent Application Award (USA), 1997 IETE-R.L. Wadhwa Gold Medal, 2001 INSA-S.H. Zaheer Medal, 2005-06 Indian Science Congress-P.C. Mahalanobis Birth Centenary Gold Medal from the Prime Minister of India for Lifetime Achievement, 2007 J.C. Bose Fellowship of the Government of India, 2013 Indian National Academy of Engineering (INAE) Chair Professorship, 2013 IETE Diamond Jubilee Medal, 2014 IEEE Fellow Class Golden Jubilee Medal, 2015 INAE-S.N. Mitra Award, 2015 DAE Raja Ramanna Fellowship, and 2017 INSA-Jawaharlal Nehru Birth Centenary Lecture award.

Prof. Pal is/ was an Associate Editor of IEEE Trans. Pattern Analysis and Machine Intelligence (2002-06), IEEE Trans. Neural Networks [1994-98 & 2003-06], Neurocomputing (1995-2005), Pattern Recognition Letters (1993-2011), Int. J. Pattern Recognition & Artificial Intelligence, Applied Intelligence, Information Sciences, Fuzzy Sets and Systems, Fundamenta Informaticae, LNCS Trans. Rough Sets, Int. J. Computational Intelligence and Applications, IET Image Processing, Ingeniería y Ciencia, and J. Intelligent Information Systems; Editor-in-Chief, Int. J. Signal Processing, Image Processing and Pattern Recognition; a Book Series Editor, Frontiers in Artificial Intelligence and Applications, IOS Press, and Statistical Science and Interdisciplinary Research, World Scientific; a Member, Executive Advisory Editorial Board, IEEE Trans. Fuzzy Systems, Int. Journal on Image and Graphics, and Int. Journal of Approximate Reasoning; and a Guest Editor of IEEE Computer, IEEE SMC and Theoretical Computer Science.