

Double Bounded Rough Set, Tension Measure, and Social Link Prediction

Suman Kundu¹ and Sankar K. Pal², *Life Fellow, IEEE*

Abstract—This 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 characterizes, 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—Granular network, link prediction, network evolution, relational data, rough set, social network, tension of relations.

I. 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 sociological creatures and they have been living in societies 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, and so on 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 the network provide a new area of research and are referred as network evolution. Several researchers have studied its different aspects. These include

Manuscript received February 12, 2018; accepted July 23, 2018. Date of publication August 20, 2018; date of current version September 11, 2018. This work was supported in part by the National Science Center, Poland, under Grant Research Project 2016/23/B/ST6/01735. (*Corresponding author: Suman Kundu.*)

S. Kundu is with the Center for Soft Computing Research, Indian Statistical Institute, Kolkata 700108, India, also with the Department of Computational Intelligence, Wrocław University of Science and Technology, 50-370 Wrocław, Poland, and also with ZINFI Software Systems Pvt. Ltd., Kolkata 700091, India (e-mail: suman@sumankundu.info).

S. K. Pal is with the Center for Soft Computing Research, Indian Statistical Institute, Kolkata 700108, India (e-mail: sankar@isical.ac.in).

Digital Object Identifier 10.1109/TCSS.2018.2861215

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], the emergence of segregation in evolving social networks [4], and user preference 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 this paper, 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. This 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 1-D 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 toward 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. Furthermore, social networks show complex overlapping-neighborhood structures, which are often indistinguishable, 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 that are in relation with either of the participating nodes 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 the granulated domain. The 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 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 that 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 that 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% nonlinked 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 as follows.

- 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 the 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 the future.

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

II. 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:

- 1) 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;
- 2) to design a link prediction algorithm as an application of the tension measure.

III. DOUBLE BOUNDED ROUGH SET

Before describing in Section IV 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., transactions in financial data, friendship in social network data, e-mail exchange in communication data, and coparticipation 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 that are related to the tuple. Granules in the domain may not be expressed with a crisp set, as there might be granules that are related to only one of the members 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 the 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 that 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 the 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.

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. Fig. 1 shows an illustration of double bounded rough set. Here, a and b represent 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 and 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 nonempty sets. For any $P \subseteq A$, there is an equivalence relation $\text{IND}(P)$ such that

$$\text{IND}(P) = \{(x, y) \in U^2 \mid \forall p \in P, p(x) = p(y)\} \quad (1)$$

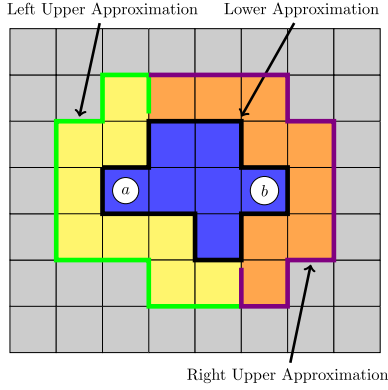


Fig. 1. Illustration: double bounded rough set.

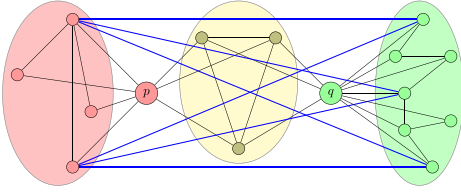


Fig. 2. Example network.

where $p(x)$ function returns the value of the attribute p for data point x . The relation $\text{IND}(P)$ is called P -indiscernibility relation and any two points $(x, y) \in P$ indicate that x and y cannot 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, and so on) 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

$$I(x) = \{[x]_P \mid \exists y \in [x]_P \text{ where } (x, y) \in Q\}. \quad (2)$$

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:

$$P_*(xQy) = I(x) \cap I(y) \quad (3)$$

$$*P(xQy) = I(x) \cup (I(x) \cap I(y)) \quad (4)$$

$$P^*(xQy) = I(y) \cup (I(x) \cap I(y)) \quad (5)$$

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

$$(*P(xQy) \cup P^*(xQy)) \setminus P_*(xQy). \quad (6)$$

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 double bounded rough set.

IV. 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 and q)] is the common neighbors. In the example network (see Fig. 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 toward each other. Our intuition says that a positive force toward 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 interneighborhood communication links. This force is also a positive one toward the relational pair (p and q). Interneighborhood links are the links between the neighbors of only p and the neighbors of only q . For example in Fig. 2, blue edges connect a node from the light red region of p 's neighbor to the 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 and 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 intraneighborhood 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 intraneighborhood 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 toward their relationship. The regions of f_3 and f_4 are shown in Fig. 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.

A. 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 $e(\cdot)$ 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 $\Gamma(p) = \{g \in \mathcal{G} | p \in g, e(g) = 1\}$.

B. 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 and q) in the notion of the aforesaid double bounded rough set (Section III) as

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

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

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

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

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

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 interneighborhood 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 (11), as shown at the bottom of this page.

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 intraneighborhood forces are quantified by the ratio of the number of granules in the left (or right) boundary region having physical link in the network to the total number of granules in the left (or right) boundary region. The mathematical formulas of f_3 and f_4 are as follows:

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

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

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

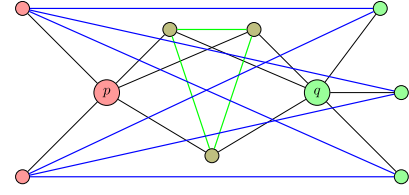


Fig. 3. Network with $\mathcal{T}(p, q) = 1$.

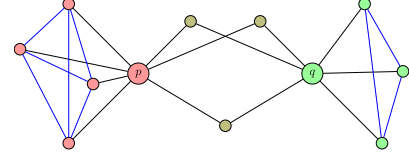


Fig. 4. Network with $\mathcal{T}(p, q) = (-1)$.

C. 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 components f_1 and f_2 contribute positively toward 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 and 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.

1) *Characteristics:* The tension value shows the following characteristics.

- 1) The value of $\mathcal{T}(p, q)$ varies from -1 to $+1$.
- 2) $\mathcal{T}(p, q) = 1$ 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 are no intraneighborhood connections. Fig. 3 shows an example network with the said situation.
- 3) 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 are no connections within the common neighbors and between interneighbors, but the intraneighborhoods of both p and q are fully connected. Fig. 4 shows a situation where $\mathcal{T} = -1$.
- 4) Tension measure along the pair (p, q) would be zero under three different conditions.

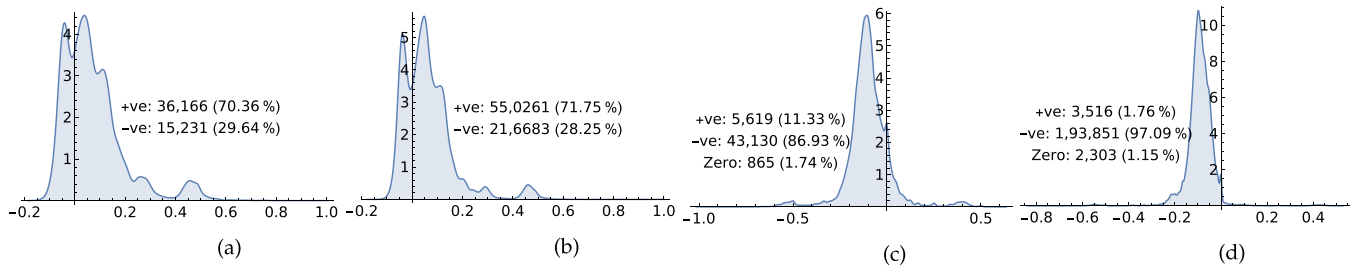


Fig. 5. Distribution of \mathcal{T} . (a) Linked pairs of data set 1. (b) Linked pairs of data set 2. (c) Nonlink pairs of data set 1. (d) Nonlink pairs of data set 2.

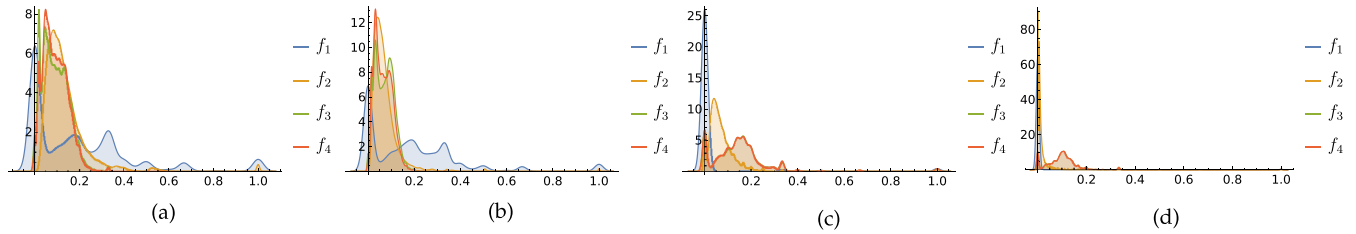


Fig. 6. Distribution of different forces. (a) Linked pairs of data set 1. (b) Linked pairs of data set 2. (c) Nonlink pairs of data set 1. (d) Nonlink pairs of data set 2.

TABLE I
SYNTHETIC DATA SETS

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

- 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.
- When $f_1 + f_2 = f_3 + f_4$, indicating a neutral strength of the relation between the pair.
- 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.

2) *Observations in Synthetically Generated Networks:* We generated two synthetic networks using LDBC [10]. The details of these data sets are shown in Table I. In each network, we calculated the tension measure for all the pairs with physical edges in the network. Fig. 5(a) and (b) shows the distributions of the value of \mathcal{T} for data sets 1 and 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. A negative value of tension for any physical link of such a network indicates that the link became inactive and it may be removed from the network in the future.

The distributions of four different forces are shown in Fig. 6(a) and (b). For both the networks, most of the

f_1 values are distributed within the range of 0–0.5, whereas the values of other three forces f_2 – f_4 are distributed between 0 and 0.2.

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

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 II. In the experiment, we examined all the edged pairs for *Wiki-Vote* [11], [12] and *Facebook* [13] networks, whereas it is 50 000 randomly selected edged pairs for *Flickr*, *Twitter*, and *Youtube* [14] data. Obtained results are shown in Table III.

It is evident from Table III 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 the negative side of the distribution indicate the possibility of their removal or deletion in the future. The distribution plots of the linked pairs for all the data sets are shown in Fig. 7.

We have also investigated the tension measure for nonexisting edged pairs. A set of 50 000 nonlinked node pairs were selected randomly from all the data sets and we measured the tension score for all these links. The distributions of

TABLE II
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 Coefficient	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 III
TENSION FOR LINKED AND UNLINKED PAIRS

Data Set	Edged Pair			Non-Edged Pair		
	+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)

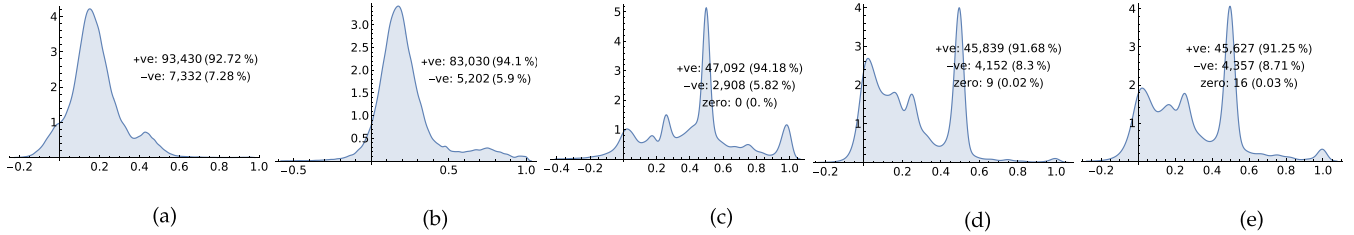


Fig. 7. Distribution of $T(p, q)$ where $e(p, q) = 1$. (a) Wiki-Vote. (b) Facebook. (c) Flickr. (d) Youtube. (e) Twitter.

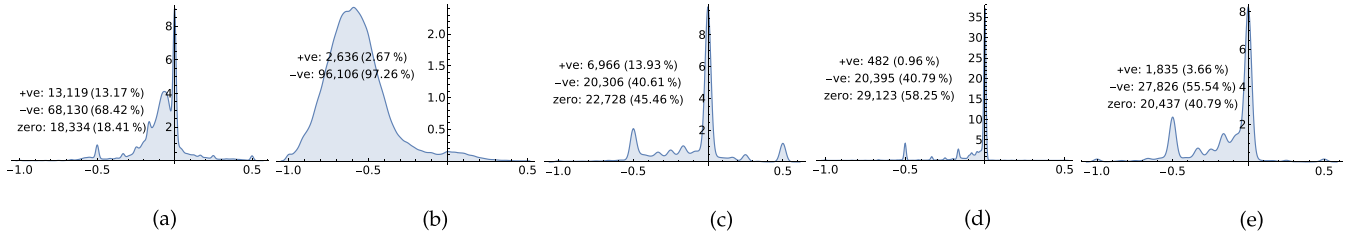


Fig. 8. Distribution of $T(p, q)$, where $e(p, q) = 0$. (a) Wiki-Vote. (b) Facebook. (c) Flickr. (d) Youtube. (e) Twitter.

tension measure for different data sets are plotted in Fig. 8. As expected, most of the tension values are either less than or equal to zero. The data summarized in Table III 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 nonlinked pairs, the tension is mostly either negative or zero. These findings validate the theoretical claims (see Section IV-C) 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 three out of five data sets?” Even for Wiki-Vote, more than 15% pairs have tension measure zero. To understand the root cause of this result, one may dig into the properties of the network. It is found that

all these data sets contain a 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 interconnection 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 (CC) of the whole network is very low. For example, here the average 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.

V. 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

Algorithm 1 Tension Calculation

```

1  '''G: networkx graph; a, b: nodes'''
2  def GetTensionValue(G, a, b):
3      ten = 0.0, c_nbrs = set([])
4      l_nbrs = [], r_nbrs = []
5
6      '''identifying nodes in the
7      rough regions of the neighborhood'''
8      tmp = set(G.neighbors(b))
9      for n in G[a]:
10         if n in tmp:
11             c_nbrs.add(n)
12         else:
13             l_nbrs.append(n)
14     r_nbrs = tmp - c_nbrs
15
16     '''Inter boundary density'''
17     f2 = 0, count = 0
18     for n in l_nbrs:
19         x = set(G[n]).intersection(r_nbrs)
20         count = count + len(x)
21
22     ncl = len(l_nbrs) * len(r_nbrs)
23     if (ncl > 0):
24         f2 = float(count)/ncl
25
26     '''common neighborhood density'''
27     count = 0, f1 = 0
28     if (len(c_nbrs) > 1):
29         cmn = set(c_nbrs)
30         while cmn:
31             node = cmn.pop()
32             x = set(G[node]).intersection(cmn)
33             count = count + len(x)
34
35         f1 = float(count)/ncr(len(c_nbrs), 2)
36
37     '''Intra neighbor density @ src'''
38     count = 0, f3 = 0
39     if (len(l_nbrs) > 1):
40         srcn = set(l_nbrs)
41         while srcn:
42             node = srcn.pop()
43             x = set(G[node]).intersection(srcn)
44             count = count + len(x)
45
46         f3 = float(count)/ncr(len(l_nbrs), 2)
47
48     '''Intra Neighbor Density @ dst end'''
49     count = 0, f4 = 0
50     if (len(r_nbrs) > 1):
51         dstn = set(r_nbrs)
52         while dstn:
53             node = dstn.pop()
54             x = set(G[node]).intersection(dstn)
55             count = count + len(x)
56         f4 = float(count)/ncr(len(r_nbrs), 2)
57
58     ten = 0.5*(f1 + f2 - f3 - f4)
59     return ten

```

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, such as 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].

A. Related Work

Having multidisciplinary 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 nonexisting 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 the 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: 1) neighbor-based metric; 2) path-based metric; and 3) 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-Newman [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, such as 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 nontopological information, such as user profile and geographical locations [36].

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

B. Algorithm Based on the Proposed Tension Measure

In this section, we propose a link prediction algorithm using the tension measure defined in Section IV-C. 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 selects the pairs with a positive tension value. Unlike the other algorithms, we do not rank the predicted links because any positive tension value (irrespective of its magnitude) refers to as a possibility of link formation between the concerned pair. The method is listed in Algorithm 2.

Algorithm 2 Link Prediction Using Tension Measure

```

1  '''input for the program is a
2  graph and list of test pairs'''
3  def GetPredictedLinks(G, test_pairs):
4      output = {}
5
6      for (x,y) in test_pairs:
7          t = GetTensionValue (G,x,y)
8          if t > 0:
9              output[(x,y)] = t
10
11     return output

```

C. Descriptions of Data Sets

In our experiments, we used two types of real-world network data: 1) without temporal information and 2) with temporal information. Nine data sets used are described below.

- 1) *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 an undirected network.
- 2) *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.
- 3) *Youtube*: The Youtube data were collected by the OSNR group of the Max Planck Institute for Software Systems and publicly available in [44]. The properties are listed in Table II.
- 4) *Flickr*: The data were collected by us in our laboratory. The data contain the Flickr user-to-user links crawled during November and December 2015. The crawl was initiated by the seeds of Indian origin in Flickr. Total link crawled was about 2 million and the features of the network are shown in Table II.
- 5) *Twitter*: The network is formed by the reply/mentions of tweets collected by us from 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 II.

- 6) *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.
- 7) *Digg* [47], [48]: The temporal data set is a friendship network of digg users collected in 2009.
- 8) *Epinion* [49], [50]: This is a temporal trust network of Epinion, an online product rating site. The network consists of individual users connected by directed trust links.
- 9) *EUEmail*: The network was generated using e-mail data from a large European research institution by the SNAP [43] group. An edge of this network represents an e-mail communication between the members.

D. Comparing Methods

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

- 1) *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

$$CN(x, y) = \Gamma(x) \cap \Gamma(y) \quad (15)$$

where $\Gamma(\cdot)$ represents the neighbors of (\cdot) .

- 2) *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 and 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

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

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

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

- 4) *Preferential Attachment*: Based on the scale free properties of social networks, it is a well-known fact that a user with many friends tends to create more future

TABLE IV
COMPARISON IN TERMS OF \mathcal{E}

	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 \mathcal{A}

connections. Based on this, the preferential attachment score is calculated as

$$\text{PA}(x, y) = |\Gamma(x) \times \Gamma(y)|. \quad (18)$$

5) *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. Jaccard's coefficient is measured by

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

E. Experimental Setup

We have conducted experiments to compare the performance of different similarity matrices [see (16)–(19)] with the proposed tension measure [see (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 nonlined pair (say E''_{prob}) of similar size. Thus, the total link pair selected as prob set is $E_{\text{prob}} = E'_{\text{prob}} \cup E''_{\text{prob}}$. 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.

1) *Effectiveness (\mathcal{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 nonrelevant objects, i.e., the false positive (fp). So, we measured the effectiveness with the difference between the true positive rate and the false positive rate as

$$\mathcal{E} = \frac{tp}{tp + fn} - \frac{fp}{tn + fp} \quad (20)$$

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

2) *Accuracy (\mathcal{A})*: Accuracy measure of a predictor is

$$\mathcal{A} = \frac{tp + tn}{tp + fp + tn + fn}. \quad (21)$$

3) *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 nonlinked pair. Considering n independent comparisons, the accuracy is measured as

$$\text{AUC} = \frac{n' + 0.5n''}{n} \quad (22)$$

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 nonlinked pair.

4) *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

$$f = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (23)$$

where precision and recall are

$$\text{Precision} = \frac{tp}{tp + fp} \quad (24)$$

$$\text{Recall} = \frac{tp}{tp + fn}. \quad (25)$$

F. Results

Table IV shows the comparative values of the measure effectiveness \mathcal{E} for different data sets. For all the data sets, tension-based link prediction algorithm shows the 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 \mathcal{E} at least 10% better than others, except the Facebook data, where tension is superior by 4.3%. A graph showing the box whiskers' plot of the obtained results is shown in Fig. 9 for the measure \mathcal{E} .

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

TABLE V
COMPARISON IN TERMS OF \mathcal{E}

	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 VI
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 VII
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

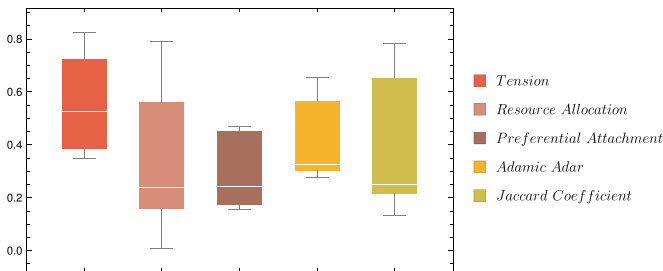


Fig. 9. Effectiveness of different data sets for different algorithms. The white bar within the box shows the median value and the upper and lower boundaries of the box are the upper and lower quartiles of the obtained values. The whiskers show the highest and lowest values reported.

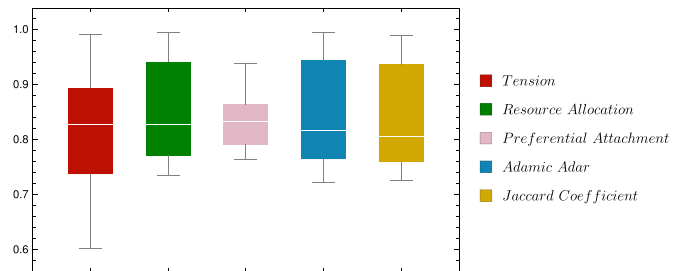


Fig. 11. AUC of different data sets for different algorithms. The white bar within the box shows the median value and the upper and lower boundaries of the box are the upper and lower quartiles of the obtained values. The whiskers show the highest and lowest values reported.

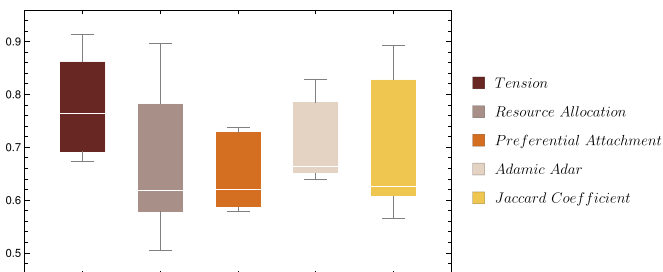


Fig. 10. Accuracy of different data sets for different algorithms. The white bar within the box shows the median value and the upper and lower boundaries of the box are the upper and lower quartiles of the obtained values. The whiskers show the highest and lowest values reported.

2.74%, and 5.28%, respectively, for Wiki-Vote, Facebook, Flickr, Youtube, and Twitter data. Fig. 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 nonlinked pairs. Table VI 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 the lowest for the tension measure as compared to the other comparing similarity matrices. A plot showing the obtained results is in Fig. 11.

Finally, the results of f -score are shown in Table VII. 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. Fig. 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 a 2.20-GHz-based system with

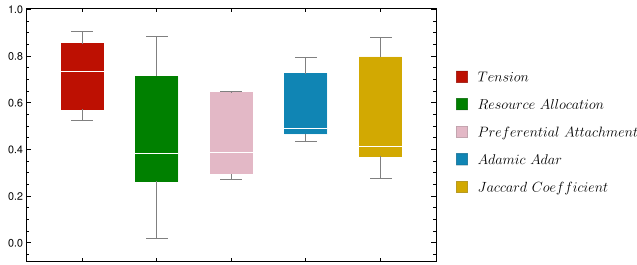


Fig. 12. f -score of different data sets for different algorithms. The white bar within the box shows the median value and the upper and lower boundaries of the box are the upper and lower quartiles of the obtained values. The whiskers show the highest and lowest values reported.

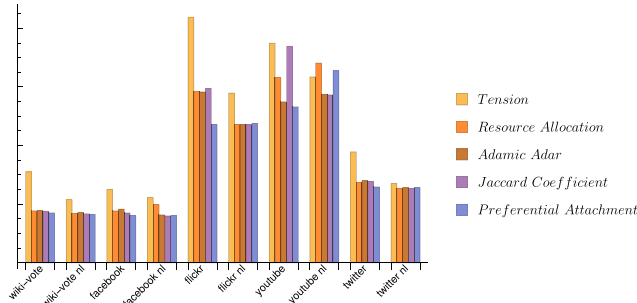


Fig. 13. Execution time in log.

8-GB 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 a few data sets as an example) are shown in Fig. 13. Each group corresponds to a particular list of test pairs. “nl” in the graph label refers to the nonlink (E''_{prob}) prob set. Execution timewise, the method based on the proposed \mathcal{T} is the slowest among 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 Fig. 13 is that for all the data sets, the execution time of the proposed method for nonlinked 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 nonlined pairs for most of the data sets.

VI. CONCLUSION AND DISCUSSION

The contribution of the current investigation is threefold. 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 variant 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 (see Section III) 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. Even 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 the participating nodes toward each other, while the negative one indicates their repulsion. Experimentally, we showed that more than 90% linked and less than 15% nonlinked pairs of nodes have a positive tension value. It is also found that for all the cases, above 85% tension value is either negative or zero for nonlinked 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.

ACKNOWLEDGMENT

The authors would like to thank A. Kundu for fruitful discussion about the problems in early stage of the research. They would also like to thank R. Das for providing them with Twitter data set used in experimental analysis. Prof. S. K. Pal would like to thank for the DAE Raja Ramanna Fellowship and the Sir J. C. Bose National Fellowship (grant-in-aid) of the Government of India.

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,” *Int. J. Behav. Develop.*, 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. 3rd Workshop Social Netw. Mining Anal. (SNA-KDD)*. Paris, France: ACM Press, 2009, pp. 9:1–9:10.
- [4] A. D. Henry, P. Prafat, and C.-Q. Zhang, “Emergence of segregation in evolving social networks,” *Proc. Nat. Acad. Sci. USA*, 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. Symp. Appl. Comput. (SAC)*. New York, NY, USA: ACM, 2017, pp. 1090–1091.
- [6] S. Kundu and S. K. Pal, “FGSN: Fuzzy granular social networks—Model and applications,” *Inf. Sci.*, vol. 314, pp. 100–117, Sep. 2015.

- [7] S. Kundu and S. K. Pal, "Fuzzy-rough community in social networks," *Pattern Recognit. Lett.*, vol. 67, pp. 145–152, Dec. 2015.
- [8] L. A. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets Syst.*, vol. 90, no. 2, pp. 111–127, 1997.
- [9] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*. Norwell, MA, USA: Kluwer, 1992.
- [10] A. Prat. *DATAGEN: Data Generation for the Social Network Benchmark*. Accessed: Mar. 11, 2015. [Online]. Available: <http://ldbouncil.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. 19th Int. Conf. World Wide Web (WWW)*. New York, USA: ACM, 2010, pp. 641–650.
- [12] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.* New York, NY, USA: ACM, 2010, pp. 1361–1370.
- [13] J. J. McAuley and J. Leskovec, "Learning to discover social circles in ego networks," in *Proc. 25th Int. Conf. Adv. Neural Inf. Process. Syst.*, P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Lake Tahoe, NV, USA: Curran Associates, 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. 7th ACM SIGCOMM Conf. Internet Meas.*, San Diego, CA, USA, 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 Trans. Knowl. Data Eng.*, 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 Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2263–2274, Oct. 2017.
- [17] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Amer. Soc. Inf. Sci. Technol.*, 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," *Sci. China Inf. Sci.*, vol. 58, pp. 1–38, Jan. 2014.
- [19] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. Lett. E*, vol. 64, no. 2, pp. 251021–251024, 2001.
- [20] A.-L. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaboration," *Phys. A*, vol. 311, nos. 3–4, pp. 590–614, 2002.
- [21] E. Ravasz, A. L. Somera, and D. A. Mongru, "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," *Soc. Netw.*, 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, 1912.
- [24] E. A. Leicht, P. Holme, and M. E. Newman, "Vertex similarity in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 73, no. 2, p. 026120, 2006.
- [25] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *Eur. Phys. J. B*, vol. 71, no. 4, pp. 623–630, 2009.
- [26] Y. X. Zhu, L. Lü, Q. M. Zhang, and T. Zhou, "Uncovering missing links with cold ends," *Phys. A, Stat. Mech. Appl.*, 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," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, 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. 27th Annu. ACM Symp. Appl. Comput. (SAC)*, Trento, Italy, 2012, pp. 138–143.
- [30] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos, "Fast and accurate link prediction in social networking systems," *J. Syst. Softw.*, vol. 85, no. 9, pp. 2119–2132, 2012.
- [31] G. Jeh and J. Widom, "SimRank: A measure of structural-context similarity," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Edmonton, AB, 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 Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 355–369, Mar. 2007.
- [33] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Washington, DC, USA, 2010, pp. 243–252.
- [34] A. De, S. Bhattacharya, S. Sarkar, N. Ganguly, and S. Chakrabarti, "Discriminative link prediction using local, community, and global signals," *IEEE Trans. Knowl. Data Eng.*, 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 Trans. Knowl. Data Eng.*, 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 Trans. Knowl. Data Eng.*, 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," in *Online Communities and Social Computing*. Berlin, Germany: Springer, 2011, pp. 336–344.
- [38] H. Kashima and N. Abe, "A parameterized probabilistic model of network evolution for supervised link prediction," in *Proc. Int. Conf. Data Mining (ICDM)*, Washington, DC, USA, 2006, pp. 340–349.
- [39] M. Pujari and R. Kanawati, "Link prediction in complex networks by supervised rank aggregation," in *Proc. IEEE 24th Int. Conf. Tools Artif. Intell. (ICTAI)*, vol. 1, Nov. 2012, pp. 782–789.
- [40] E. Tasnádi and G. Berend, "Supervised prediction of social network links using implicit sources of information," in *Proc. 24th Int. Conf. World Wide Web (WWW)*, 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. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Chicago, IL, USA, 2013, pp. 775–783.
- [42] M. Kaya, M. Jawed, E. Bütün, and R. Alhaji, "Unsupervised link prediction based on time frames in weighted-directed citation networks," in *Trends in Social Network Analysis (Lecture Notes in Social Networks)*, R. Missaoui, T. Abdesslem, and M. Latapy, Eds. Cham, Switzerland: Springer, 2017.
- [43] J. Leskovec. *SNAP Large Network Data*. Accessed: Aug. 21, 2017. [Online]. Available: <https://snap.stanford.edu/data/index.html>
- [44] *Online Social Networks Research at The Max Planck Institute of Software Systems*. Accessed: Aug. 23, 2017. [Online]. Available: <http://socialnetworks.mpi-sws.org/data-ipc2007.html>
- [45] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *IEEE Trans. Mobile Comput.*, vol. 6, no. 6, pp. 606–620, Jun. 2007.
- [46] (Apr. 2017). *Haggle Network Dataset—KONECT*. [Online]. Available: <http://konect.uni-koblenz.de/networks/contact>
- [47] T. Hogg and K. Lerman, "Social dynamics of Digg," *EPJ Data Sci.*, vol. 1, p. 5, Jun. 2012.
- [48] (Apr. 2017). *Digg Friends Network Dataset—KONECT*. [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. Amer. Assoc. Artif. Intell. Conf.*, 2005, pp. 121–126.
- [50] (Apr. 2017). *Epinions Trust Network Dataset—KONECT*. [Online]. Available: <http://konect.uni-koblenz.de/networks/epinions>
- [51] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
- [52] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.



Suman Kundu received the B.Tech. degree in information technology from the West Bengal University of Technology, Kolkata, India, in 2005, and the M.E. degree in software engineering and the Ph.D. degree in engineering from Jadavpur University, Kolkata, in 2009 and 2017, respectively.

He was a Senior Research Fellow with the Center for Soft Computing Research, Indian Statistical Institute, Kolkata, from 2010 to 2015. He was with software development company ZINFI Software Systems Pvt. Ltd., Kolkata, from 2006 to 2010 and 2016 to 2018. He is currently a Post-Doctoral Researcher with the Wrocław University of Science and Technology, Wrocław, Poland. He published four journal and two conference articles in social network analysis, granular computing, soft computing, and distributed computing. For more details, visit website <http://sumankundu.info>.



Sankar K. Pal (M'80–SM'84–F'93–LF'15) received the Ph.D. degree in radio physics and electronics from the University of Calcutta, Kolkata, India, in 1979, and the Ph.D. degree in electrical engineering along with DIC from Imperial College London, London, U.K., in 1982.

He was with the University of California at Berkeley, Berkeley, CA, USA, and the University of Maryland at College Park, College Park, MD, USA, from 1986 to 1987; the NASA Johnson Space Center, Houston, TX, USA, from 1990 to

1992 and in 1994; and the U.S. Naval Research Laboratory, Washington, DC, USA, in 2004. He has held several visiting positions in Italy, Poland, Hong Kong, and Australian universities. He is a Distinguished Scientist and the former Director of the Indian Statistical Institute, Kolkata, and a former Chair Professor of the Indian National Academy of Engineering, New Delhi, India. He is currently a Distinguished Professor Chair of the Indian National Science Academy (INSA), New Delhi. He founded the Machine Intelligence Unit and the Center for Soft Computing Research: a National Facility in the Institute in Kolkata. He has co-authored 20 books and over 400 research publications in 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 and teaching in India.

Dr. Pal is a fellow of The World Academy of Sciences, the International Association for Pattern recognition, the International Association of Fuzzy Systems, the International Rough Set Society, and all the four National Academies for Science/Engineering in India. He is/was a member of the Executive Advisory Editorial Board of the IEEE TRANSACTIONS ON FUZZY SYSTEMS, the *International Journal on Image and Graphics*, and the *International Journal of Approximate Reasoning*. He received the 1990 S. S. Bhatnagar Prize (which is the most coveted award for a scientist in India), the 2013 Padma XShri Award (one of the highest civilian awards)

from the President of India, and many prestigious awards in India and abroad, including the 1999 G. D. Birla Award, the 1998 Om Bhasin Award, the 1993 Jawaharlal Nehru Fellowship, the 2000 Khwarizmi International Award from the President of Iran, the 2000–2001 FICCI Award, the 1993 Vikram Sarabhai Research Award, the 1993 NASA Tech Brief Award, USA, the 1994 IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS Outstanding Paper Award, the 1995 NASA Patent Application Award, USA, the 1997 IETE-R. L. Wadhwa Gold Medal, the 2001 INSA-S. H. Zaheer Medal, the 2005–2006 Indian Science Congress-P. C. Mahalanobis Birth Centenary Gold Medal from the Prime Minister of India for Lifetime Achievement, the 2007 J. C. Bose Fellowship of the Government of India, the 2013 Indian National Academy of Engineering (INAE) Chair Professorship, the 2013 IETE Diamond Jubilee Medal, the 2014 IEEE Fellow Class Golden Jubilee Medal, the 2015 INAE-S. N. Mitra Award, the 2015 DAE Raja Ramanna Fellowship, and the 2017 INSA-Jawaharlal Nehru Birth Centenary Lecture Award. He is/was an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE from 2002 to 2006, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS from 1994 to 1998 and from 2003 to 2006, *Neurocomputing* from 1995 to 2005, *Pattern Recognition Letters* from 1993 to 2011, the *International Journal of Pattern Recognition and Artificial Intelligence*, *Applied Intelligence*, *Information Sciences*, *Fuzzy Sets and Systems*, *Fundamenta Informaticae*, *LNCS Transactions on Rough Sets*, the *International Journal of Computational Intelligence and Applications*, the *IET Image Processing*, *Ingeniería y Ciencia*, and the *Journal of Intelligent Information Systems*; an Editor-in-Chief of the *International Journal of Signal Processing, Image Processing and Pattern Recognition*; a Book Series Editor of *Frontiers in Artificial Intelligence and Applications* (IOS Press) and *Statistical Science and Interdisciplinary Research* (World Scientific); and a Guest Editor of the IEEE *Computer*, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, and *Theoretical Computer Science*. He visited 45 countries as a Keynote/Invited Speaker or an Academic Visitor. Since 1997, he has been serving as a Distinguished Visitor of the IEEE Computer Society, USA, for the Asia-Pacific Region.