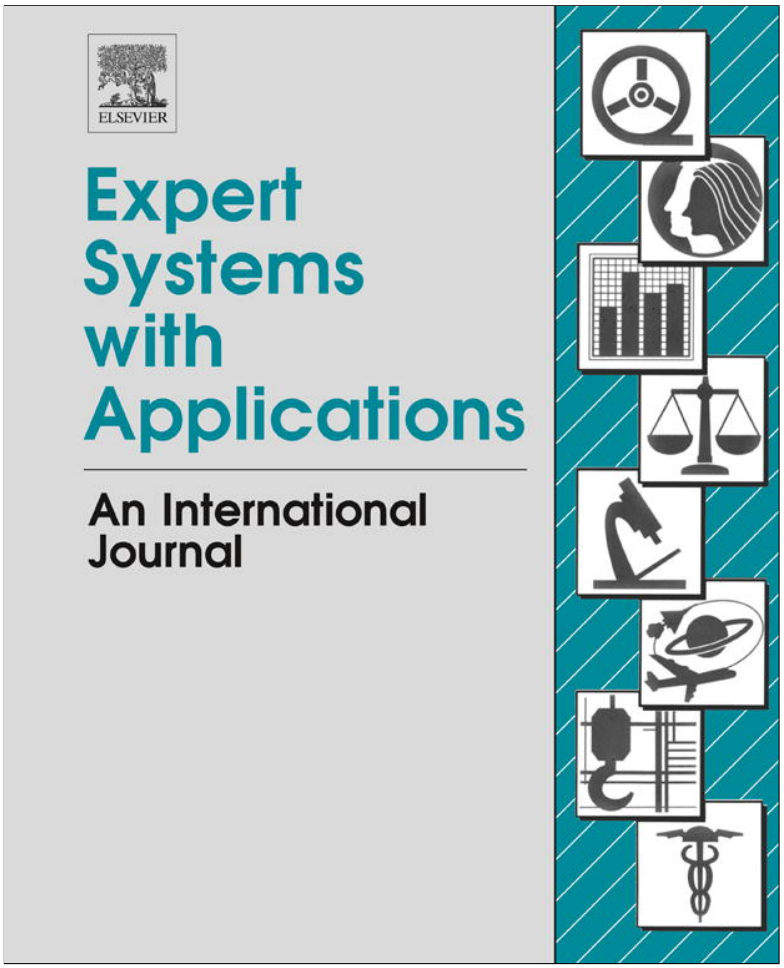

Expert Systems with Applications 40 (2013) 6652–6660

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Proximity measures for link prediction based on temporal events

Paulo R.S. Soares, Ricardo B.C. Prudêncio *

Center of Informatics (CIn), Federal University of Pernambuco, Recife, PE, Brazil

ARTICLE INFO

Keywords:
Link prediction
Temporal events
Neighborhood-based measures
Co-authorship networks

ABSTRACT

Link prediction is a well-known task from the Social Network Analysis field that deals with the occurrence of connections in a network. It consists of using the network structure up to a given time in order to predict the appearance of links in a close future. The majority of previous work in link prediction is focused on the application of proximity measures (e.g., path distance, common neighbors) to non-connected pairs of nodes at present time in order to predict new connections in the future. New links can be predicted for instance by ordering the pairs of nodes according to their proximity scores. A limitation usually observed in previous work is that only the current state of the network is used to compute the proximity scores, without taking any temporal information into account (i.e., a static graph representation is adopted). In this work, we propose a new proximity measure for link prediction based on the concept of temporal events. In our work, we defined a temporal event related to a pair of nodes according to the creation, maintenance or interruption of the relationship between the nodes in consecutive periods of time. We proposed an event-based score which is updated along time by rewarding the temporal events observed between the pair of nodes under analysis and their neighborhood. The assigned rewards depend on the type of temporal event observed (e.g., if a link is conserved along time, a positive reward is assigned). Hence, the dynamics of links as the network evolves is used to update representative scores to pairs of nodes, rewarding pairs which formed or preserved a link and penalizing the ones that are no longer connected. In the performed experiments, we evaluated the proposed event-based measure in different scenarios for link prediction using co-authorship networks. Promising results were observed when the proposed measure was compared to both static proximity measures and a time series approach (a more competitive method) that also deploys temporal information for link prediction.

1. Introduction

As the amount of interaction among individuals increases in virtual environments, more useful social data become available, serving as a basis for social network analysis. Social networks are structures composed by individuals (entities) that can be connected by different forms of social relationship (such as friendship, in which two individuals are connected if they are friends) (Amaral, Scala, Barthelemy, & Stanley, 2000). In social networks, connections and entities tend to appear and disappear along time, which turns them into highly dynamic and complex systems. Social Network Analysis (SNA) is a broad field of research that tries to deal with such complexity (Wasserman & Faust, 1994). Several tasks can be associated to SNA. In this paper, our specific aim is to investigate the prediction of links in a social network, that is, we are focused on predicting the most probable future connections based on previous states of the network. This is a well-known problem from the SNA field called *link prediction* (Hasan, Chaoji, Salem, & Zaki, 2006).

A lot of work has been devoted to cope with the link prediction problem (Getoor & Diehl, 2005; Wang, Satuluri, & Parthasarathy, 2007; Xiang, 2008). The majority of previous work is based on the application of proximity measures to non-connected pairs of nodes in the network at present time in order to predict new connections at future time. The proximity measures are used to associate scores to the pairs of nodes, which can be used either: (1) in a unsupervised approach, in which a chosen score is simply ordered and links are predicted for the top ranked pairs; or (2) in a supervised approach, in which the link prediction is treated as a classification task and different scores are used as predictor attributes by a learning algorithm. A limitation that can be pointed out in the previous work is that the proximity scores are usually calculated without taking into account the evolution of the network. The proximity measures are computed using all network data up to the present moment (i.e., the current network state) without considering when links were created. Hence, a potentially useful source of information for link prediction is not adequately leveraged.

* Corresponding author. Tel.: +55 8121268430.
E-mail addresses: prss@cin.ufpe.br (P.R.S. Soares), rbcp@cin.ufpe.br (R.B.C. Prudêncio).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.06.016

In the current work, we propose a novel proximity measure in which temporal information is taken into account. In the proposed measure, we deployed the concept of *temporal events*, which are specific activities (e.g., creation or removal of a link) observed between a pair of nodes in consecutive time intervals. For instance, an *innovative* event occurs when two nodes are not connected in a given time interval and a new link is created between them in the next interval. The proximity score for a pair of nodes is computed by monitoring along time the events observed around the nodes and their immediate neighbors. Each category of temporal event defined in our work (innovative, conservative and regressive) is associated to a numerical reward and the proximity score increases or decreases along time depending on the temporal events observed in consecutive network frames. Our proposal was supported by the Homans' work that associates the strength of a connection between individuals to their frequency of interaction (which was modeled in our work by the temporal events) (Homans, 1951) and the idea shared by many authors that a bigger common neighborhood between individuals is closely related to a higher probability of future connections (Newman, 2001).

In order to verify the viability of the proposed measure, we performed experiments on co-authorship networks extracted from four sections of the physics e-Print arXiv.[1] In these experiments, the proposed measure was evaluated in different scenarios by considering for instance different reward values for the temporal events and by adopting a weighting function to give higher importance to more recent events. For a baseline comparison, experiments were performed with traditional proximity measures previously adopted in the literature. We also performed experiments with a time series approach (Potgieter, April, Cooke, & Osunmakinde, 2009) which, similarly to our approach, also aims to consider temporal information to improve link prediction. In all experiments, the unsupervised approach was adopted to perform the prediction task once the scores were calculated and the Area Under the ROC Curve (AUC) was used to evaluate the performance of prediction. The performance achieved by the event-based measure outperformed the results of the baseline methods in all the networks considered.

Section 2 briefly presents the link prediction problem. Section 3 describes the event-based approach, showing the concepts used and the score calculation process. Section 4 presents the social network data adopted, the experiments and obtained results. Finally, Section 5 concludes this work by presenting some final considerations and future work.

2. Link prediction

Link prediction consists of predicting new connections or detecting hidden links in a network. It is a very important task applicable to a wide variety of areas, such as bibliographic domain, molecule biology, criminal investigations and recommending systems (Getoor & Diehl, 2005; Xiang, 2008). A traditional definition of the link prediction problem is expressed by: "Given a snapshot of a social network at time t, we seek to accurately predict the edges that will be added to the network during the interval from time t to a given future time t'" (Liben-Nowell et al., 2003). Among several approaches to treat the problem, the most widespread ones rely on the use of *proximity* measures between pairs of nodes (Lu & Zhou, 2011; Xiang, 2008). As previously mentioned, the starting point of this approach is to extract the values/scores of different measures that indicate the similarity between pairs of nodes. These scores can be used either by unsupervised (Liben-Nowell et al., 2003; Lu & Zhou, 2011; Murata & Moriyasu, 2008) or supervised link prediction (Hasan et al., 2006; Lichtenwalter, Lussier, & Chawla, 2010; Sá & Prudêncio, 2011). In the former approach, a proximity measure is chosen and deployed to rank node pairs in the network. The top ranked ones are predicted to be linked. In the supervised link prediction, a set of proximity measures is chosen and adopted as predictor attributes by a classifier, learned based on historical data. Each pair of non-connected nodes is described by the set of proximity scores. A classifier then uses these attributes to perform a binary classification to decide whether the link will be formed or not in the future.

The proximity measures proposed and evaluated in the literature can be broadly categorized into *semantic* or *topological* measures (Xiang, 2008). In the semantic measures (or node-wise measures), the nodes' content is considered to measure proximity. For instance, in a co-authorship network, the similarity between keywords extracted from published papers can be used to predict future interaction among the authors (Xiang (2008)). Different from the semantic measures, the topological strategy consists of deploying the network structure to compute the proximity scores (e.g. the number of common neighbors that two nodes share). Topological measures are more commonly adopted in the literature since they are more general and do not require the definition of rich features to describe content (in fact, rich content is not always available depending on the social network considered).

Several topological measures were proposed in the literature mainly categorized into *neighborhood-based* or *path-based* measures (Hasan & Zaki, 2011). The neighborhood-based measures take into account the immediate neighbors of the nodes. In general, these measures consider that two nodes are more likely to form a link if their sets of neighbors have a large overlap (Xiang, 2008). Among the neighborhood-based measures, we can mention Common Neighbors (Newman, 2001), Preferential Attachment (Barabasi et al., 2002; Newman, 2001),Adamic and Adar (2003) and the Jaccard's coefficient (Salton & McGill, 1986). The path-based measures in turn define proximity between nodes by considering the paths between them. The basic idea is that two nodes are more likely to form a link if there are short paths between them. The path-based measures range from the simple path-distance measure to more sophisticated definitions that consider ensembles of different paths, such as the Katz measure (Katz, 1953). In comparative terms, the neighborhood-based methods are more widespread, due to both their computational eficiency and great performance observed in experiments (Huan, 2006; Liben-Nowell et al., 2003; Murata & Moriyasu, 2008). The measure proposed in our work can be categorized as neighborhood-based, since it uses information about the connections around nodes to assign scores to them (see Section 3).

Most previous work performs link prediction by statically analyzing the network data, that is, no temporal information is considered to perform the prediction. However, temporal information (e.g., the moments when two nodes interacted in the past or the time when a connection was first observed) is an important aspect that should be considered during the link prediction (Hasan & Zaki, 2011). For instance, when computing a proximity score based on neighborhood, it would be interesting to consider not only *how many* but also *when* the links were formed between neighbors. Recent activity between common neighbors may be more important than older activity. Also, static approaches are appropriate to investigate whether a certain link will ever occur in a network but they are less useful, for instance, to applications for which the prediction of repeated link occurrences are of interest (Huang & Lin, 2009).

In this section, we mention some previous work concerning the use of temporal information for link prediction. For instance, in Tylenda, Angelova, and Bedathur (2009), the authors modeled a network as a weighted graph, in which the weight of a link was the age of the most recent activity between the corresponding

[1] http://www.arxiv.org

nodes. The link prediction task was then accomplished by deploying extended versions of the proximity measures adequate for weighted networks (e.g., weighted Adamic Adar). Bringmann, Berlingerio, Bonchi, and Gionis (2010) proposed to mine network data augmented with temporal information in order to discover association rules (in terms of frequent subgraphs) that best explained the network evolution. In Juszczyszyn, Musial, and Budka (2011), the authors adopted a related approach in which the history of the network (recorded during past time windows) is used to derive probabilities of transitions between triads of nodes. Although promising results can be obtained in this approach, mining frequent subgraphs has shown to be a very expensive task (Coenen, Jiang, & Zito, 2012).

An alternative approach for deploying temporal information is to treat link prediction as a time series forecasting problem (Huang & Lin, 2009; Potgieter et al., 2009; Qiu, He, & Yen, 2011; Soares & Prudêncio, 2012). Huang and Lin (2009) built a time series for each pair of nodes, in which each series observation is the frequency of occurrence of links between the nodes during a specific time period. The time series forecasts produced by ARIMA models (Box & Jenkins, 1970) were then used to estimate the probability of future link occurrence. In Potgieter et al. (2009) and Soares and Prudêncio (2012), the authors adopted a similar idea, but in this case time series models were used to predict proximity scores. In this approach, given a chosen proximity measure, a time series is built for each pair of nodes by computing the score in a sequence of time periods (i.e., using different snapshots of the network along time). A final proximity score is then obtained for each pair by forecasting its corresponding time series. Different models were applied in the forecasting process, including ARIMA models, smoothing methods and linear regression models. Despite the good results obtained in experiments, a difficulty in the time series approach is to choose an adequate forecasting model, among the variety of models that can be applied. In fact, in Soares and Prudêncio (2012), the authors observed that a wrong decision concerning the forecasting model can have a negative impact in the link prediction performance.

3. Event-based link prediction

Social networks are highly dynamic structures in which several connections and nodes tend to appear or disappear along time. The temporal evolution of these networks brings valuable information about how connections tend to be formed, and, for that, should be considered in the link prediction task. In the current work, we propose a new proximity measure that takes into account the temporal structure of a network and temporal events related to pairs of nodes in the network. The aim of our proposal is to combine Homans' thought that the strength of a connection between two individuals is directly associated with how often they interact with one another (modeled here by means of temporal events) (Homans, 1951) and Newman's idea that the bigger the number of common neighbors between two nodes, the higher is their probability to be connected in the future (Newman, 2001).

The general idea of the proposed approach is to increase or decrease the proximity score between two nodes depending on *temporal events* observed among the two nodes and their neighborhood. A temporal event, which will be explained better in the next section, is defined through the appearance or disappearance of links around nodes as the network evolves. In the proposed work, initially a temporal structure is created by extracting consecutive *frames* of the network at different time intervals. A proximity score is then computed for each pair of nodes by aggregating the rewards assigned for the temporal events observed during the transitions of frames. The proposed measure is explained in more detail in the next subsections.

3.1. Temporal structure

In the proposed solution, we initially build a temporal structure $\mathcal{N}$, by following a methodology described in Soares and Prudêncio (2012). First of all, the network is split into several time-sliced snapshots, which represent states of the network at different time intervals in the past. After that, *frames* are built by grouping consecutive snapshots. The size of each frame is the same as the length of the *prediction window*, pre-defined in the link prediction task. Each frame represents one time step in $\mathcal{N}$. This methodology is detailed below.

Let $G(V,E)$ be a graph representing a social network observed up to time T. Each edge in E is represented by a triple (u,v,t), indicating that the nodes u and $v \in V$ had a social interaction at time t. Hence, more than one edge related to a single pair of nodes can be observed in E if the nodes interacted in different moments in the past. Let w be the length of the prediction window, i.e., our task is to predict new links concerning the future time interval from $T+1$ to $T+w$.

Let G_t be the sub-graph of G containing only the edges observed at time t. Let $[G_t, G_{t+1}, \ldots, G_{t+m}]$ be the frame formed by the disjoint union of the graphs from time t to $t+m$. In our work, a set of n consecutive frames $\mathcal{N} = \{F_1, \ldots, F_k, \ldots, F_n\}$ of size w is extracted from the graph G. Formally, $\mathcal{N}$ can be defined as

$$\begin{aligned}\mathcal{N} = \{ \quad & [G_{T-nw+1}, G_{T-nw+2}, \ldots, G_{T-(n-1)w}], \ldots, \\ & [G_{T-2w+1}, G_{T-2w+2}, \ldots, G_{T-w}], \\ & [G_{T-w+1}, G_{T-w+2}, \ldots, G_T]\}\end{aligned} \tag{1}$$

The frame $F_k = [G_{T-(n-k+1)w+1}, \ldots, G_{T-(n-k)w}]$ is hence the sub-graph of G containing all links observed in the k-th time interval of size w. As an example, consider a network observed up to the year $T = 2012$ and a prediction window of length 2 (i.e., the task is to predict new links from 2013 and 2014). If we extract $n = 3$ frames from the network, the following structure ($\mathcal{N} = \{F_1, F_2, F_3\}$) is obtained

$$\mathcal{N} = \{[G_{2007}, G_{2008}], [G_{2009}, G_{2010}], [G_{2011}, G_{2012}]\} \tag{2}$$

This structure can be used to analyze how the network evolved every 2 years since 2007. For this, we will introduce next section the concept of *temporal events*, that can be observed between subsequent frames.

3.2. Temporal events

We define a *temporal event* as a specific activity between two nodes from a frame to its subsequent. A temporal event is the action that leads a dyad (a pair of nodes) from a state (connected or non-connected) to another. Events can be categorized into one of three mutually exclusive types: *conservative*, *innovative* or *regressive*, defined below.

- *Conservative*
  A conservative event occurs when a relationship between two nodes is not dropped when the network evolves, that is, when two nodes share a link in a frame and this connection is preserved in the subsequent one. For each pair of nodes (u,v), we define a reward $\mathcal{C}(u,v,k)$ related to frame F_k, in order to take into account a conservative event during the transition from the $(k-1)$th to the kth frame. Formally:

$$\mathcal{C}(u,v,k) = \begin{cases} c, & \text{if } (u,v) \in E_{k-1} \cap E_k \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

In the above definition, E_{k-1} and E_k are the sets of edges observed in frames F_{k-1} and F_k respectively. The constant c indicates the reward for conservative events, which should be a non-negative value since the strength of a tie between two nodes is preserved.

- *Innovative*
  Innovative events represent the creation of a new link between two nodes on different frames. They happen when two nodes are not connected in a frame and a link is observed in the next frame. The Innovative reward $\mathcal{I}(u,v,k)$ associated to a pair (u,v) and a frame F_k is:

$$\mathcal{I}(u,v,k) = \begin{cases} i, & \text{if } (u,v) \in E_k \setminus E_{k-1} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

The constant i in the above equation indicates the reward for innovative events. Its value should be positive since the tie between two nodes is strengthened.

- *Regressive*
  Regressive events are opposite to innovative ones. This kind of event represents the removal of an existing link between two nodes from a frame to its subsequent. The Regressive reward $\mathcal{R}(u,v,k)$ is defined as:

$$\mathcal{R}(u,v,k) = \begin{cases} r, & \text{if } (u,v) \in E_{k-1} \setminus E_k \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

In this event, r should assume a non-positive value since the strength of the connection between the nodes tends to decrease.

The values of the parameters c, i and r can be determined empirically by evaluating the link prediction performance in a validation set and choosing the best configuration of values considered. As it will be seen, in our experiments, the number of parameters to define was reduced to two by making the values of c and r proportional to i (in this case, i was set to 1 for the sake of simplicity).

As it will be seen in the next subsection, given a pair of nodes, we will define as *primary* events the ones strictly related to both nodes in the pair, whereas *secondary* events happen in dyads composed by only one of the nodes in the pair and its neighbors.

3.3. Event-based score

As previously said, many approaches for link prediction compute scores to pairs of nodes by deploying a chosen proximity measure (see Section 2), aiming to determine how similar those nodes are and, consequently, how likely a connection between them will be formed in a close future.

The proposed measure combines both: (1) the rewards associated to *primary events*, which are the temporal events strictly related to the pair of nodes under analysis; and (2) the rewards associated to *secondary events*, which are the temporal events observed in the nodes' neighborhood. The proximity score associated to a given pair of nodes (u,v) is defined as:

$$score(u,v) = \sum_{k=2}^{n} P(u,v,k) + \alpha.S(u,v,k) \quad (6)$$

$$P(u,v,k) = \mathcal{C}(u,v,k) + \mathcal{I}(u,v,k) + \mathcal{R}(u,v,k) \quad (7)$$

$$S(u,v,k) = \sum_{y \in \Gamma(u) \cap \Gamma(v)} P(u,y,k) + P(y,v,k) \quad (8)$$

In Eq. (7), $P(u,v,k)$ computes the reward of the event (conservative, innovative or regressive) for the pair of nodes (u,v) observed in the transition from frame $k-1$ to frame k. In Eq. (8), $S(u,v,k)$ indicates the aggregated reward of secondary events associated to the pair (u,v) (i.e., the primary events observed in its common neighborhood). In this equation, $\Gamma(x)$ is the set of neighbors of the node x in the network.

In Eq. (6), the parameter α is an amortization factor that indicates how strong secondary events affect the tie between u and v. $P(u,v,k)$ and $S(u,v,k)$ associated to the first frame (i.e., $k=1$) are all null since this frame is not a result of any set of events, and hence for simplicity, they are not considered in Eq. (6).

Fig. 1 illustrates the application of the proposed event-based score. Suppose that one wants to compute the proximity score for the pair $(1,3)$. The nodes 1 and 3 have node 2 as common neighbor in the network. Hence, the score of the connection between 1 and 3 is going to be calculated as a function of the events that occurred between themselves (primary events), plus the ones occurred in the dyads $(1,2)$ and $(2,3)$ (secondary events) along the network temporal structure $\mathcal{N}$.

From frame F_1 to frame F_2, it can be noticed that a conservative and a regressive event happened in the dyad $(1,2)$ and $(2,3)$ respectively. So far, no event occurred in the dyad $(1,3)$. Therefore, up to frame F_2, the partial score of the pair $(1,3)$ is given by the combination of the amortized rewards associated with the secondary events described above, that is, $\alpha(c+r)$.

Looking at the next time step (frame F_3), a regressive event is associated to $(1,2)$, whereas an innovative one occurred in the pair $(2,3)$. In this step, there is also an innovative event related to the pair under analysis $(1,3)$; its presence must be counted too in the final score of the pair (its reward, however, will not be amortized since it is a primary event). Hence, the resultant score in this frame is given by $\alpha(r+i)+i$. By the definition of Eq. (6), the scores are cumulative in time. The final score of a pair of nodes is, then, the sum of all its partial scores along $\mathcal{N}$. This, therefore, results in a final score of $\alpha(c+2r+i)+i$ for the pair $(1,3)$.

In the current work, we also propose an extension of the event-based score that increases the importance of more recent events observed for a pair of nodes. In Eq. (9), we propose a proximity score which considers a monotonically increasing function $\beta : \mathbb{Z} \rightarrow \mathbb{R}$ to weight recent events more heavily than the old ones. In our work, we adopted a logarithmic function (see Eq. (10)), although other functions can be deployed. We highlight that Eq. (6) is a special case of Eq. (9) by setting $\beta(k) = 1$.

$$score(u,v) = \sum_{k=2}^{n} \beta(k).[P(u,v,k) + \alpha.S(u,v,k)] \quad (9)$$

$$\beta(k) = log(k) \quad (10)$$

3.4. Prediction

Finally, in order to perform the prediction of new links, we adopted an unsupervised strategy after the scores were computed for the pairs of nodes. It basically consists of ranking the pairs of nodes according to their scores, and then, selecting the top-ranked ones as the new links in the network, as explained in Section 2.

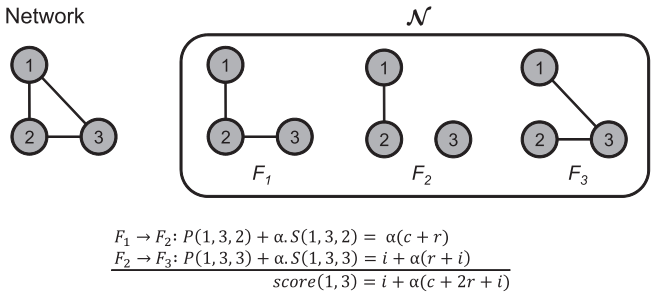

Fig. 1. Example of event-based score computation using a network composed by a click of three nodes. $\mathcal{N}$ represents the temporal structure of the network divided in three frames (F_1, F_2 and F_3). The variables c, i and r are the rewards associated with conservative, innovative and regressive events respectively, and α is the amortization factor used to deal with secondary events.

4. Experiments and results

In this section we describe the experiments performed to evaluate the proposed approach as well as the obtained results. Initially we describe the social network data used and the methodology of experiments (Section 4.1), followed by the baseline methods adopted for a comparative evaluation (Section 4.2). In Section 4.3, we present the initial results aimed to evaluate the proposed measure considering different configurations of rewards for the temporal events. In Section 4.4, we evaluate the influence of the amortization factor α in the proposed measure. Finally, in Section 4.5, we present the effect of using a weighting function for recent event as well as the comparative results with the baseline methods.

4.1. Data and settings

In order to evaluate the performance of the proposed scores, data from four co-authorship networks were used in our experiments. We highlight that co-authorship networks are highly used in the literature to evaluate SNA techniques. Being a social structure, this kind of network is very dynamic along time, what makes it a great source of data to be explored regarding the emergence of new connections. Besides, most of them are publicly available in digital environment (Barabasi et al., 2002; Newman, 2001). In a co-authorship network, a node represents an author and an edge indicates that two specific nodes have co-authored a paper. As seen in Section 3, in our work an edge also stores the time when the relationship was observed. More specifically, in the co-authorship networks adopted here, each edge stores the publication year of the co-authored paper.

The network data used in our experiments were collected from the e-print arXiv,[2] which maintains a large database of electronic scientific papers in several fields, such as mathematics, physics, astronomy, among others. In our experiments, the networks were built using collaborations from four distinct sub-areas, namely: astro-physics (astro-ph), condensed matter (cond-mat), high energy physics – lattice (hep-lat) and theoretical high energy physics (hep-th). Collaborations from 1993 to 2000 were used for astro-ph and cond-mat, and collaborations from 1992 to 2010 were used for hep-lat and hep-th. The summarized information about the adopted networks are presented in Table 1.

In our experiments, a prediction window of size one was adopted, and hence the aim of the analysis is to investigate how the networks evolved every year. By following the methodology described in Section 3.1, a temporal structure $\mathcal{N}$ was built for each network. For astro-ph and cond-mat, data from 1993 to 1999 were used to build $\mathcal{N}$ (with a total number of $n = 7$ frames) and the snapshot from 2000 was used as the prediction frame. For hep-lat and hep-th, $\mathcal{N}$ was built with collaborations from 1992 to 2009 (with a total number of $n = 18$ frames), whereas the prediction frame was composed by collaborations from 2010.

As mentioned in Section 3.2, adequate values for the reward parameters can be estimated by evaluating the link prediction performance on a validation set. In our work, the last frame of $\mathcal{N}$ was adopted as the validation set in order to perform parameter selection for each network: the frame of 1999 was used to validation for the astro-ph and cond-mat networks and the frame of 2009 was used to validation for hep-lat and hep-th.

As usual in the link prediction field, our aim is to investigate the prediction of new links in a network. Therefore, for each network, the initial list of candidate pairs to be evaluated is formed by pairs that are not connected in the last frame of $\mathcal{N}$. The final list is formed by picking (from the initial list) only the dyads that were directly or indirectly affected by some event during the network evolution. Table 2 shows the distribution of the candidate pairs of nodes regarding their class labels (connected or non-connected) both in the validation and the test sets.

[2] arXiv.org e-Print archive – Cornell University Library.

Table 1
Networks statistics.

	astro-ph	cond-mat	hep-lat	hep-th
Papers	19,077	20,664	9,367	38,569
Authors	16,978	18,070	4,718	17,887
Collaborations	140,157	56,731	32,309	58,855
Density (10^{-4})	9.7252	3.4746	29.0355	3.6792
Avg. degree	16.51	6.28	13.70	6.58

As it can be seen, the class distribution is highly imbalanced, which is a common problem related to the link prediction task. In order to minimize the effect of imbalanced data during evaluation, the Area Under ROC Curve (AUC) was adopted as the measure of performance. ROC curves relate the sensitivity (true positive rate) and specificity (true negative rate) of a classifier. The AUC has been traditionally used in imbalanced classification problems (Murata & Moriyasu, 2008; Wang et al., 2007) due to its higher robustness to class distribution anomalies (Fawcett, 2006). Finally, for more reliable results, we performed a bootstrap over the test set by taking 100 randomly stratified sub-samples and computing the AUC. The 10% of the best results and 10% of the worst results were discarded and the average AUC was computed.

4.2. Baseline methods

In our experiments, the proposed measures were compared with different measures previously adopted in the literature of link prediction. Initially, four classical proximity scores based on neighborhood were considered: PA, CN, AA and JC. These are very popular measures used in the state of the art of link prediction (Bringmann et al., 2010; Lichtenwalter et al., 2010). Each measure is explained below.

The PA measure assumes that the probability of a future link between two nodes is proportional to their degrees. In a co-authorship network, such probability is correlated to the product of the number of collaborators they have Barabasi et al. (2002). Hence, it is defined as:

$$PA(u, v) = ||\Gamma(u)|| \times ||\Gamma(v)|| \quad (11)$$

The CN measure states that the bigger the number of neighbors two nodes share, the higher is their probability to form a link in the future (Newman, 2001). Formally, the measure is defined as:

$$CN(u, v) = ||\Gamma(u) \cap \Gamma(v)|| \quad (12)$$

The AA measure refines the CN by increasing the scores of pairs of nodes in which the neighbors in common possess less connections (Adamic & Adar, 2003). Its formal definition is:

Table 2
Class distribution for the validation and test sets.

	astro-ph	cond-mat	hep-lat	hep-th
(a) Validation set				
No. of pairs	745,683	149,769	330,137	498,782
Connected (+)	7,797	1,737	1,099	1,232
Non-connected (−)	737,886	148,032	329,038	497,550
Proportion ($^{+}/_{-}$)	1.06%	1.17%	0.33%	0.25%
(b) Test set				
No. of pairs	1,337.607	254,651	373,053	553,544
Connected (+)	9,791	2,665	608	1,626
Non-connected (−)	1,327.816	251,986	372,445	551,918
Proportion ($^{+}/_{-}$)	0.74%	1.06%	0.24%	0.29%

$$AA(u,v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{log(||\Gamma(z)||)} \quad (13)$$

Finally, the JC assumes higher proximity values for pairs of nodes which share a higher *proportion* of common neighbors relative to total number of neighbors they have Salton and McGill (1986).

$$JC(u,v) = \frac{||\Gamma(u) \cap \Gamma(v)||}{||\Gamma(u) \cup \Gamma(v)||} \quad (14)$$

The above measures were statically applied to the whole graph G (i.e., in the current state of the network) without considering temporal information. In our work, we also performed experiments with the time series approach for link prediction (Potgieter et al., 2009; Soares & Prudêncio, 2012) described in Section 2. As our current work, this approach also considers temporal information during the link prediction. Initially the network is split into frames, using the same procedure described in Section 3.1. Given a chosen proximity measure, a time series is built for each pair of nodes by applying the measure in each network frame. A forecasting model is then used in order to predict the next value of the series. Such forecast value is defined as the final proximity score of the pairs of nodes. Compared to the proposed work, we can point out a difficulty related to the time series approach, which is the choice of a good forecasting model. In fact, in the experiments performed in Soares and Prudêncio (2012) using co-authorship networks, the choice of the forecasting model strongly affected the link prediction performance. In our work, we performed experiments adopting AA as the measure to build the time series and the linear regression (LR) model as the forecasting model. That was the best combination of proximity measure and forecasting model in the experiments presented in Soares and Prudêncio (2012).

Finally, we highlight that similarly to the proposed measures, the baseline measures described in this section (both the static ones and the time series approach) were applied for unsupervised link prediction.

4.3. First-round experiments

As described in Section 3.3, the event-based measure is potentially affected by the rewards associated to each of the three events: conservative, innovative and regressive. In the first round of experiments, we aimed: (1) to evaluate the effect of rewards' values in the performance of the proposed measure; and (2) to verify whether these values can be properly chosen using a validation set. We deployed the validation set to empirically evaluate and select the best combination of rewards in the link prediction task. In this round, as mentioned in Section 3.2, i was fixed to 1.0 and the other two parameters c and r varied proportionally to $i : c = \{0.0, 0.25, 0.5, 1.0, 2.0\}$ and $r = \{-2.0, -1.0, -0.5, -0.25\}$. We consider here a fixed value for α, since our focus was to investigate the adopted rewards' values. The value of α was set to 0.05 through a preliminary and exploratory battery of experiments. Table 3 shows the AUC value obtained by each combination of the rewards in the validation set for all networks. The best four observed results for each network are presented in bold.

As it can be seen, the most effective combinations of reward did not vary so much from one network to another. The best results were obtained by parameters (c and r) around the same region (lower central) in the tables for all networks. Besides, these results indicate that conservative and regressive rewards can balance each other. The best performance values were achieved when $c \leqslant i$, although some good results were also obtained when we considered a higher weight (2.0) to the conservative reward. In our experiments, if regressive reward is roughly weighted, the event-based method did not perform well (this can be noticed by analyzing results obtained with $r = -2.0$ and -1.0). As one could expect, this shows that if the aim is to predict new links in a network, then the history of innovative and conservative events assumes a higher degree of importance in the prediction process.

Table 3
AUCs obtained by different combinations of rewards on the validation sets.

r/c	0.0	0.25	0.5	1.0	2.0
(a) astro-ph					
−2.0	0.3402	0.3552	0.3698	0.4074	0.4681
−1.0	0.6301	0.6657	0.6773	0.6919	0.6972
−0.5	0.7931	0.7998	0.8012	0.8018	0.7996
−0.25	0.7967	**0.8021**	**0.8035**	**0.8041**	**0.8023**
(b) cond-mat					
−2.0	0.3002	0.3087	0.3181	0.3530	0.4239
−1.0	0.5684	0.6202	0.6318	0.6483	0.6507
−0.5	0.7928	0.8083	**0.8131**	**0.8128**	0.8092
−0.25	0.7964	0.8086	**0.8109**	**0.8123**	0.8105
(c) hep-lat					
−2.0	0.3792	0.4553	0.5801	0.6849	0.7273
−1.0	0.8002	0.8330	0.8294	0.8206	0.8102
−0.5	**0.8406**	**0.8476**	**0.8467**	**0.8424**	0.8340
−0.25	0.8309	0.8374	0.8389	0.8378	0.8331
(d) hep-th					
−2.0	0.2610	0.2843	0.3276	0.4264	0.5223
−1.0	0.6668	0.7599	0.76554	0.7678	0.7584
−0.5	**0.8501**	**0.8579**	**0.8555**	**0.8502**	0.8409
−0.25	0.8433	0.8497	0.8491	0.8469	0.8411

Table 4 shows the AUC value obtained by each combination of the rewards in the test set for all networks. The results observed in the test sets did not varied substantially from those ones observed in the validation set. The best configurations of parameters were also similar considering the validation and the test set (see Table 5). This supports the conclusions we made regarding rewards and indicates that they can be determined empirically using a validation set.

4.4. Second-round experiments

In the second round of experiments, our aim was to investigate how the amortization factor α affects the performance of the event-based measure. As explained in Section 3.3, this parameter indicates the importance of the secondary events in relation to primary events in our proposed measure. In this experiment, we eval-

Table 4
AUCs obtained for several combinations of c and r on the test sets.

r/c	0.0	0.25	0.5	1.0	2.0
(a) astro-ph					
−2.0	0.3462	0.3684	0.3916	0.4401	0.5116
−1.0	0.6423	0.6781	0.6875	0.6979	0.7006
−0.5	0.7844	0.7911	**0.7922**	0.7916	0.7879
−0.25	0.7861	**0.7919**	**0.7927**	**0.7929**	0.7909
(b) cond-mat					
−2.0	0.3041	0.3109	0.3259	0.3743	0.4303
−1.0	0.5731	0.6325	0.6408	0.6544	0.6535
−0.5	0.8072	**0.8154**	**0.8147**	0.8122	0.8064
−0.25	0.8078	**0.8126**	**0.8127**	0.8123	0.8091
(c) hep-lat					
−2.0	0.2945	0.3618	0.4934	0.6431	0.7353
−1.0	0.7652	0.8480	0.8429	0.8403	0.8349
−0.5	0.8483	**0.8573**	**0.8574**	**0.8557**	0.8515
−0.25	0.8393	0.8478	0.8506	**0.8527**	0.8508
(d) hep-th					
−2.0	0.2258	0.2376	0.2706	0.3583	0.4634
−1.0	0.6624	0.7303	0.7331	0.7303	0.7220
−0.5	**0.8605**	**0.8625**	**0.8587**	0.8529	0.8441
−0.25	0.8555	**0.8558**	0.8541	0.8518	0.8466

Table 5
Best rewards settings extracted from experiments on validation and test sets.

	astro-ph	cond-mat	hep-lat	hep-th
(a) Validation set				
c	1.0	0.5	0.25	0.25
i	1.0	1.0	1.0	1.0
r	−0.25	−0.5	−0.5	−0.5
(b) Test set				
c	1.0	0.25	0.5	0.25
i	1.0	1.0	1.0	1.0
r	−0.25	−0.5	−0.5	−0.5

uated the proposed measure by adopting different values of α ranging from 0 to 1.5, with an increment of 0.05. We highlight that when $\alpha = 0$, secondary events are in fact not taken into account. For higher values (such as $\alpha = 1.5$) in turn, possibly an excessive importance is given to secondary events, which may harm the prediction performance. In this section, the rewards c, r and i were fixed according to the best configuration observed in the validation set in the previous round of experiments.

The AUC values obtained as α varies are illustrated in Fig. 2. When α assumes a null value, the performance of the model is harmed because secondary events are not being taking into account in the prediction task. In this case, the measure will only be able to predict connections between pairs affected by primary events any time in the past, that is, pairs that were connected in at least one frame of $\mathcal{N}$. This shows the effectiveness of secondary events in the prediction of new connections.

The best AUC values were obtained by adopting $\alpha = 0.05$, which is a relatively small value considering the range of values evaluated in our experiments. When it approaches to higher values, the method performs worse since surrounding pairs are becoming as important as the main pair of nodes in analysis. Besides, the number of secondary events is much bigger than the total amount of primary events (mainly if the network is dense). Hence, isolated secondary events should contribute with small values, which when aggregated, can outstand in the predictive process without, however, mask the effects of the primary events. In this section, we suggest a simple heuristic to define the value of α based on the ratio between the total number of primary events and secondary events in $\mathcal{N}$ (see Table 6). As it can be seen, the AUC results obtained by adopting the heuristic value for α were similar to the best results observed in Fig. 2.

4.5. Last-round experiments

In the previous experiments we evaluated the event-based score focusing on its parameters. In this section, we evaluated it aiming to determine if the age of the temporal events (event-based score with Eq. (9)) can bring any gain of performance to the prediction task. Additionally, we compared the event-based scores to the baseline measures described in Section 4.2. Table 7 present the

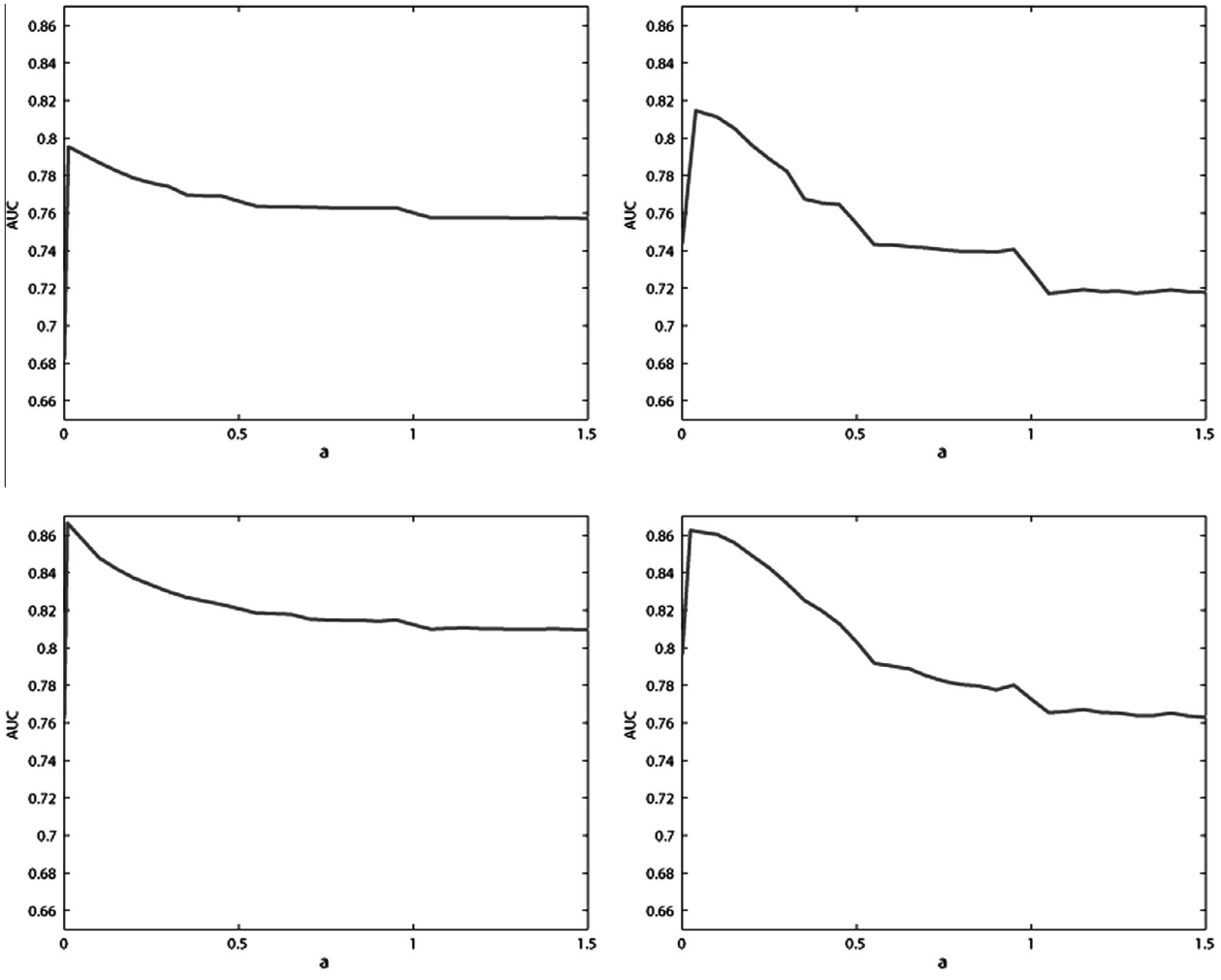

Fig. 2. Plot of AUC versus the amortization factor α.

Table 6
Proportion between the number of primary events (P) and secondary events (S).

	astro-ph	cond-mat	hep-lat	hep-th
$^{P}/_{S}$	0.0116	0.0391	0.0089	0.0238
AUC with $\alpha = P/S$	0.7953	0.8160	0.8664	0.8627
Best AUC	0.7948	0.8269	0.8681	0.8830

Table 7
Performance of the methods (AUCs).

	astro-ph	cond-mat	hep-lat	hep-th
Event-based ($\beta = log$)	0.7948	0.8269	0.8681	0.8830
Event-based ($\beta = 1$)	0.7929	0.8147	0.8573	0.8625
Time series (AA + LR)	0.7584	0.7402	0.8537	0.7899
AA	0.7844	0.7346	0.8049	0.7513
CN	0.7391	0.6744	0.7598	0.6838
JC	0.7299	0.6114	0.7408	0.6481
PA	0.5043	0.5455	0.5668	0.4834

AUC value obtained by the proposed scores and the baseline methods. In this table, we refer the function assigned to β to distinguish between the scores generated by using the two aforementioned equations.

The results showed that the methods that deploy temporal information (i.e., the event-based scores and the time series approach) in general outperformed the classical static approach based on PA, CN, AA and JC. These results confirm that temporal information is actually an important aspect to consider for link prediction. The event-based scores obtained the best results from all methods considered. The performance gain was higher for the cond-mat, hep-lat and hep-th networks, but less expressive for the astro-ph network, in which the event-based score was quite similar to AA alone. In our results, we observed a performance gain when the event-based score using $\beta = log$ was compared to the event-based score using $\beta = 1$ in all networks considered, what supports our belief that recent events carry more information about the emergence of links.. Although, the performance gain was in general small in absolute values, it was statistically verified using a t-test with a 95% level of confidence.

Finally, although both the time series approach and the event-based methods explore the temporal nature of the network, the proposed approach obtained the best comparative results. The performance gain was statistically verified with a 95% level of confidence. The proposed measure analyses the network by focusing on the connections as such, and for that, it was able to extract more relevant information to the prediction of ties that might emerge in a close future. The method based on time series, on the other hand, assumes that the topological measure history shows some tendency as the network evolves. Thus, it performs an indirect analysis of the connections in the network. In most cases, the time series approach outperformed the static approach, but it was not able to overcome the results achieved by an analysis strictly focused on the links, as done by the method based on events.

5. Conclusion

In this work we introduced a new proximity measure for link prediction based on the notion of temporal events. Differently from the classical static approach that takes the network at time t in order to predict new connections at future time t', our method takes temporal information into account by monitoring how the network evolved along time. Our method consists of computing scores by aggregating rewards of the temporal events observed at each step in the network evolution, which resulted in more representative scores to the dyads under analysis. Different experiments were performed in four co-authorship networks, initially aimed to evaluate the robustness of the proposed method concerning its parameters. The obtained results suggest that the method was quite stable concerning the definition of its parameters across the different networks adopted, although more experiments need to be performed in other domains of application. Also, we verified the importance of considering secondary events in the proposed measure.

In the performed experiments, we also compared the event-based score to other baseline approaches proposed in the literature of link prediction. The obtained results showed that the event-based measure outperformed the baseline measures considered, indicating that the combination of interactions between nodes and the dynamics of common neighborhood can bring a gain in performance to the prediction task. Additionally, when the age of the temporal events was taken into account, the method performed slightly better, indicating that more recent events can bring more useful information to the prediction process. In our work, we adopted a *log* function in order to give a higher weight to more recent events. Nevertheless, other functions can be evaluated in the future (e.g., linear) to weight the recent events. Also, different functions can be associated to each category of temporal event if we assume for instance that recent events from one category are actually more important that the other events.

Finally, we highlight that the event-based score described here was specific to undirected and unweighted networks. A possible future line of research is to extend the ideas discussed here to more complex categories of networks. For instance, a variety of temporal events could be defined if the direction of the links is considered (e.g., a conservative event is observed in a direction but a regressive event is observed in the other direction, thus indicating a non-reciprocal interaction between the nodes at a given moment). Also, in the case of weighted networks, temporal events could consider an increase or decrease of the weight between the nodes along time, instead of just considering the existence of the links. The rewards in such cases should be carefully defined in order to adequately reflect the important of the events in the link prediction process.

References

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social Networks, 25*(3), 211–230.

Amaral, L. A., Scala, A., Barthelemy, M., & Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences USA, 97*(21), 11149–11152.

Barabasi, A-L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A, 311*(3–4), 590–614.

Box, G., & Jenkins, G. (1970). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Bringmann, B., Berlingerio, M., Bonchi, F., & Gionis, A. (2010). Learning and predicting the evolution of social networks. *Intelligent Systems IEEE, 25*(4), 26–35.

de Sá, H. R., & Prudêncio, R. B. C. (2011). Supervised link prediction in weighted networks. In *Proceedings of the 2011 international joint conference on neural networks* (pp. 2281–2288). IEEE.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters, 27*, 861–874.

Getoor, L., & Diehl, C. P. (2005). Link mining: A survey. *SIGKDD Explorations Newsletter, 7*(2), 3–12.

Hasan, M. A., Chaoji, V., Salem, S., Zaki, M. (2006). Link prediction using supervised learning. In *Proceedings of SDM 06 workshop on link analysis counterterrorism and security*.

Hasan, M. A., & Zaki, M. J. (2011). A survey of link prediction in social networks. In Charu C. Aggarwal (Ed.), *Social network data analytics* (pp. 243–275). US: Springer.

Homans, G. C. (1951). *The human group*. London: Routledge and Kegan.

Zan Huan, (2006). Link prediction based on graph topology: The predictive value of the generalized clustering coefficient. In *Proceedings of 12th ACM SIGKDD international conference on knowledge discovery and data mining.*

Huang, Z., & Lin, Dennis K. J. (2009). The time-series link prediction problem with applications in communication surveillance. *INFORMS Journal on Computing, 21*(2), 286–303.

Jiang, C., Coenen, F., & Zito, M. (2013). A survey of frequent subgraph mining algorithms. *The Knowledge Engineering Review, 28*(1), 75–105.

Juszczyszyn, K., Musial, K., Budka, M. (2011). Link prediction based on subgraph evolution in dynamic social networks, In *Privacy security risk and trust (PASSAR) 2011 IEEE third international conference on social computing* (pp. 27–34).

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika, 18*(1), 39–43.

Liben-Nowell, D., Kleinberg, J. (2003). The link prediction problem for social networks. In *Proceedings of the 2003 international conference on information and knowledge management* (pp. 556–559).

Lichtenwalter, R. N., Lussier, J. T., Chawla, N. V. (2010). New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 243–252).

Lu, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A, 390*(6), 1150–1170.

Murata, T., & Moriyasu, S. (2008). Link prediction based on structural properties of online social networks. *New Generation Computing, 26*(3), 245–257.

Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks. *Physical Review E, 64*.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. In *Proceedings of the national academy of sciences of the United States of America* (Vol. 98, pp. 404–409).

Potgieter, A., April, K. A., Cooke, R. J. E., & Osunmakinde, I. O. (2009). Temporality in link prediction: Understanding social complexity. *Journal of Emergence: Complexity and Organization, 11*(1), 83–96.

Qiu, B., He, Q., Yen, J. (2011). Evolution of node behavior in link prediction. In *AAAI* (pp. 1810–1811).

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. McGraw-Hill Inc.

Soares, P. R. S., Prudêncio, R. B. C. (2012). Time series based link prediction. In *Proceedings of the 2012 international joint conference on neural networks* (pp. 784–790).

Tylenda, T., Angelova, R., Bedathur, S. (2009). Towards time-aware link prediction in evolving social networks. In *Proceedings of the third workshop on social network mining and analysis, SNA-KDD '09* (pp. 1–9).

Wang, C., Satuluri, V., Parthasarathy, S. (2007). Local probabilistic models for link prediction. In *Proceedings of the 2007 seventh IEEE international conference on data mining* (pp. 322–331).

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications (structural analysis in the social sciences)*. Cambridge University Press.

Xiang, E. W., (2008). A survey on link prediction models for social network data. PhD thesis, PhD Qualifying Exam, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology.