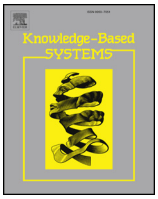
Knowledge-Based Systems 156 (2018) 129–137

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



# Combining contextual, temporal and topological information for unsupervised link prediction in social networks

Carlos Pedro Muniz, Ronaldo Goldschmidt[*], Ricardo Choren

*Military Institute of Engineering, Computer Engineering Department, Praca General Tiburcio 80, Rio de Janeiro, Brazil*

ARTICLE INFO

*Keywords:*
Link prediction
Social network analysis
Data mining
Graph mining
Temporal and contextual information

*2017 MSC:*
00-01
99-00

ABSTRACT

Understanding and characterizing the processes driving social interactions is one of the fundamental problems in social network analysis. In this context, link prediction aims to foretell whether two not linked nodes in a network will connect in the near future. Several studies proposed to solve link prediction compute compatibility degrees as link weights between connected nodes and, based on a weighted graph, apply weighted similarity functions to non-connected nodes in order to identify potential new links. The weighting criteria used by those studies were based exclusively on information about the existing topology (network structure). Nevertheless, such approach leads to poor incorporation of other aspects of the social networks, such as context (node and link attributes), and temporal information (chronological interaction data). Hence, in this paper, we propose three weighting criteria that combine contextual, temporal and topological information in order to improve results in link prediction. We evaluated the proposed weighting criteria with two popular weighted similarity functions (Adamic-Adar and Common Neighbors) in ten networks frequently used in experiments with link prediction. Results with the proposed criteria were statistically better than the ones obtained from the weighting criterion that is exclusively based on topological information.

## 1. Introduction

With the growth of the Internet and the creation of social media, people and organizations can engage, interact, and communicate more easily, establishing a large social network that allows for knowledge sharing in virtual environments [1]. A social network is a highly interconnected graph in which nodes represent participants and links represent one or more types of interdependence (association) amid corresponding participants. It grows and changes quickly over time through the addition of new links, denoting the appearance of new interactions in the social structure [2].

Understanding and characterizing the processes driving social interactions is one of the fundamental problems in social network research [3]. In this scenario, link prediction appears as a central problem of social network analysis, aiming to infer which unobserved links will appear in the near future by a given snapshot of a network [4]. Link prediction has many important applications, e.g., recommending friends in an online social network or inferring whether two authors will establish coauthor relationship in a bibliographic network [5].

Several studies have been proposed to predict links in social networks [6–10]. In general these studies fall into two main approaches [7,11,12]: supervised and unsupervised. In the supervised approach, the original graph that represents the social network is converted to a binary classification problem. Then, algorithms such as decision trees or probabilistic methods are used to build classification models. The unsupervised approach (also known as similarity-based approach [10]) uses similarity functions ($d$: $V \times V \to \mathbb{R}$) that compute scores to express some sort of similarity degree between pairs of nodes. A similarity degree is a numeric value used to concisely describe properties shared by two nodes. In this sense, nodes are considered similar if they share common properties observed in the network, e.g. common friends in a social network. Recently, various similarity functions to explore different properties observed in social networks have been proposed in the literature [6–10].

Although link prediction state-of-the-art reveals no common sense about which approach performs better, the supervised approach presents two important drawbacks, when compared to the unsupervised alternative: (a) it usually demands high computational complexity to build the classification models, and (b) once social networks evolve fast, their classification models may become obsolete as time goes by, requiring model replacement by new and updated versions. Hence, in this paper, we focus on the unsupervised approach to link prediction.

Most unsupervised link prediction studies compute the similarity degrees between pairs of non-connected nodes [2,4,6,13–18]. They

[*] Corresponding author.
*E-mail addresses:* cptullio@gmail.com (C.P. Muniz), ronaldo.rgold@ime.eb.br (R. Goldschmidt), choren@ime.eb.br (R. Choren).

https://doi.org/10.1016/j.knosys.2018.05.027
Received 22 November 2017; Received in revised form 19 April 2018; Accepted 19 May 2018
Available online 21 May 2018
0950-7051/ © 2018 Elsevier B.V. All rights reserved.

consider that the pair of non-connected nodes that presents the higher score is the best suited (has more probability) to connect in the future. Other studies consider the similarity between connected nodes [16,18–22], also called *link strength* (i.e. the strength of an existing connection), to provide useful information to predict new links. For example, two non-connected nodes strongly linked to their common neighbors are more likely to connect than the ones weakly linked to their common neighborhood. Algebraically, these studies consider each link strength as a numerical weight assigned to the corresponding edge of the network graph.

Data used for similarity computation can be classified in three major groups: topological, contextual and temporal. The *topological group* encloses the sort of information that is inherently related to the network structure (or topology). This kind of data is not explicitly available in the network: it must be calculated from the network structure; hence, it does not depend on the type of the analyzed network. Examples of topological data include number of common neighbors, Adamic-Adar coefficient, preferential attachment and many others. This kind of feature has been intensively investigated in research on link prediction [2,11–13,17,19–21,23,24].

The *contextual group* encompasses the sort of information that describes the attributes about the social network application domain. The existence (availability) of such information is directly dependent of the analyzed network and the set of available contextual data vary from a network to another since it is context dependent. Examples of contextual data include client's gender, job location, paper's keywords, number of co-authored papers, product review, product category, and others. Several studies on link prediction evaluated this kind of attribute [5,14,15,25–27].

The *temporal group* encloses the sort of information about important chronological data related to both topological and contextual aspects of the social network. Since temporal data traverse both aspects, it is relevant to deal with this kind of data separately. Examples of temporal data include time-stamps associated to network modifications, such as insertions of nodes and edges (temporal data related to topological aspect), paper publication year (temporal data related to contextual aspect), and so on. Recent research has evaluated the influence of the temporal data on link prediction [16–18,21,23,28].

Link strength computation in most unsupervised link prediction studies is based exclusively on topological data. They focus on the frequency of existing interactions between nodes as weighting criterion. This means that the link strength between nodes that have more interactions is higher than the link strength between nodes with less interactions. Such restriction leads to poor incorporation of other aspects of the social network. For instance, according to the homophily social theory [10], actors with similar interests generally connect in common communities to interact; for this reason, connections between nodes with similar profiles (described by contextual attributes) should have a higher link strength. The same is valid for temporal information. Time unawareness means that old and new interactions have the same influence in weight calculation. Yet, according to the Weak Ties theory [29], recent interactions tend to stimulate new interactions in the network and thus should have higher influence in link prediction. Thus, we are interested in examining if the combination of topological, contextual and temporal information can improve unsupervised link prediction. Our question is hence: *given a snapshot of an homogeneous attributed multigraph*[1] $G_\tau = \langle V, E \rangle$, *where V is the set of nodes and E is the set of links, is it still possible to enhance unsupervised link prediction by combining topological, contextual and temporal information in weight calculation*?

In this work, we initially propose a general weighting model that allows the user to configure different weighting criteria based on combinations of contextual, temporal and topological aspects observed in a social network. Then, we configure the proposed model generating three specific weighting criteria. The first criterion, Temporal-Topological (*TT*), combines the frequency of interactions between connected nodes (topological data) and the age of the most recent interaction (temporal data). The Contextual-Topological criterion (*CT*) merges the similarity between the profiles of connected nodes (contextual data) and the frequency of interactions between those nodes (temporal data). At last, the Contextual-Temporal-Topological criterion (*CTT*) gathers frequency and age of interactions with similarity between the profiles of connected nodes. *CTT* is the article's main contribution. It is a new weighting criterion that combines topological, temporal and contextual information *simultaneously*. Experimental results with ten social networks provide statistical evidence that *CTT* does enhance unsupervised link prediction when compared to other weighting criteria that do not combine the three aspects, including the current state-of-the-art criterion that is exclusively based on topological data.

The remainder of the paper is organized as follows. Some background concepts on link prediction and related work are discussed in Section 2. We present the proposed general weighting model and specific weighting criteria in Section 3. In Section 4, we conduct an experimental study to evaluate the criteria proposed. Section 5 concludes the work and presents alternatives of future initiatives.

## 2. Background

### 2.1. Link prediction

Link prediction is a basic computational problem underlying network evolution and it can be formally stated as follows. Given an homogeneous attributed multigraph $G_\tau(V, E)$, where $V$ is the set of nodes, $E \subseteq V \times V$ is the set of undirected attributed (with at least one temporal information) edges (links) and $\tau$ a time-stamp, the goal of link prediction is to foretell whether there will appear a link $e \in E$ between arbitrary, non-connected nodes $u$ and $v$ (i.e. $e = (u, v)$) at a next time-stamp $\tau + 1$.

The basic process for an unsupervised link prediction method $P$ follows the sequence of tasks first proposed by Liben-Nowell and Kleinberg [2] and it did not consider link strength. So it was later altered to accommodate a graph weighting activity for link strength computation. The modified process can be seen in Fig. 1 and it comprises the following activities:

**Activity 1: Graph Partition**. In this activity, $G$ is divided into a training ($G_{Trn}$) and a test ($G_{Tst}$) sub-graphs. The training sub-graph contains all edges created up to a given time-stamp $\tau$ and the test sub-graph encloses all edges that are present in $G$ after time-stamp $\tau$. $E_{Old}$ represents the set of edges in $G_{Trn}$ and $E_{New}$ denotes the set of edges that are in $G_{Tst}$ but were not in $G_{Trn}$. In other words, $E_{New}$ indicates the new interactions we are seeking to predict.

**Activity 2: Core Set Identification**. This activity identifies the `Core` set of nodes. It encloses those nodes that are considered active, i.e. nodes that frequently interacted with others before and after $\tau$. As social networks grow through the addition of nodes as well as edges, it is not reasonable to seek predictions for edges whose endpoints are not present in $G_{Trn}$ [2]. Thus the `Core` set is defined to be all nodes incident to at least $k_{Trn}$ edges in $G_{Trn}$ and at least $k_{Tst}$ edges in $G_{Tst}$. Parameters $k_{Trn}$ and $k_{Tst}$ are defined by the user and they typically depend on the average frequency of interactions in the network.

**Activity 3: Graph Weighting**. This activity is used to weight the edges of the social network graph so that the link prediction method can use link strength. In this activity, artificial edges are created to represent pairs of nodes that are connected in $G_{Trn}$. Then, the weight (link strength) is calculated for each artificial edge, using a weighting criterion. Weights are calculated using similarity functions such as:

[1] A graph $G$ is called: (a) homogeneous iff $G$ has one type of node and one type of edge; (b) multigraph iff $G$ contains two or more edges between two nodes; and (c) attributed iff $G$ has attributes in its nodes and/or edges.

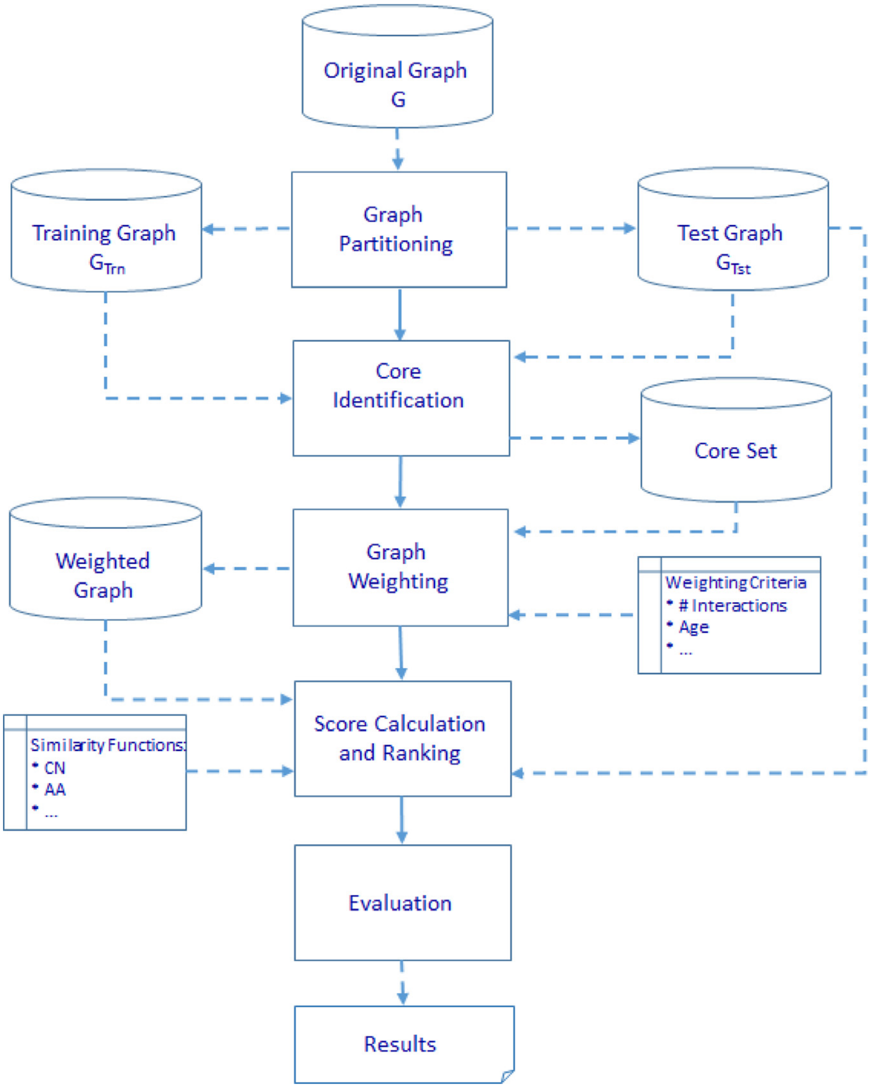


**Fig. 1.** Process for weighting-based link prediction: overview.

(A) *Frequency of Interactions* [16,18,19]: computes the number of existing interactions between two arbitrary nodes $u$ and $v$. That is: $|E(u, v)|$

(B) *Age of the Most Recent Interaction* [17,23,30]: computes the elapsed time from the last interaction between two nodes and the current time (CTime). That is: $CTime - max(t_{(u,v)})$

(C) *Age of the Oldest Interaction* [17,23,30]: computes the elapsed time from the first interaction between two nodes and the current time. That is: $CTime - min(t_{(u,v)})$

(D) *Cosine Similarity (Salton Index)* [10,13]: computes the similarity between the attributes (characteristics) of two nodes $u$ and $v$. Its calculation uses an aggregation function ($f$) that outputs the set of characteristics that describe a given node [7]. That is: $\frac{|f(u)\cap f(v)|}{\sqrt{|f(u)|\times|f(v)|}}$

**Activity 4: Score Calculation and Ranking**. This activity is executed to produce a ranked list ($L_P$) in descending order of *score* between pairs of non-connected nodes $u$ and $v$ ($u, v \in$ `Core`). Score calculation is done by similarity functions. There are several weighted similarity functions (usually weighted versions of the non-weighted similarity functions) such as:

(A) *Weighted Common Neighbors – WCN(u, v)*. It is a weighted variant of the Common Neighbors similarity function. It computes the average link strength between two given nodes $u$ and $v$ and their common neighbors. The higher the strength of the relationship between them, the bigger is the chance that $u$ and $v$ will connect. That is:

$$\sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{w(u, z) + w(z, v)}{2}$$

(B) *Weighted Adamic-Adar – WAA(u, v)*. It is a weighted variant of the Adamic-Adar similarity function. It quantifies neighborhood overlap between two given nodes $u$ and $v$. Different from *WCN*, in this function, common neighbors with smaller degrees (i.e. common neighbors with fewer neighbors) and stronger connections with $u$ and $v$ are weighted more heavily when evaluating the chances of $u$ and $v$ getting connected. That is:

$$\sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{w(u, z) + w(z, v)}{2} \times \frac{1}{log\left(\sum_{z'\in\Gamma(z)} w(z', z)\right)}$$

(C) *Weighted Preferential Attachment – WPA(u, v)*. It is a weighted variant of the Preferential Attachment similarity function. It indicates that new links will be more likely to connect nodes with stronger relationships than the ones with weaker associations. That is:

$$\sum_{u' \in \Gamma(u)} w(u, u') \times \sum_{v' \in \Gamma(v)} w(v, v')$$

**Activity 5: Evaluation**. This activity evaluates the link prediction method $P$. In this step, we take the ranked list $L_P$ of pairs and select the top $n$ pairs with the highest likelihood to connect at a time-stamp posterior to $\tau$. The value of $n$ is defined as:

$$n = |E_{New} \cap (Core \times Core)|$$

Then, the performance of $P$ is compared to the performance of a baseline random link predictor $P_{rand}$ which simply randomly selects pairs of nodes that did not interact in the training interval. A random prediction is correct with probability expressed as:

$$C_{P_{rand}} = \frac{|E_{New}|}{\binom{Core}{2} - |E_{Old}|}$$

The improvement factor of $P$ over random is calculated as (where $|E_{Correct}|$ is the number of links correctly predicted by the link predictor):

$$ImpFactor_P = \frac{|E_{Correct}|/|E_{New}|}{C_{P_{rand}}}$$

### 2.2. Related work

In the following, we give an overview of related work that followed the weighting-based procedure described in Fig. 1. Mostly, studies from this group differ in the weighting criterion, i.e. the way similarity functions for link strength calculation are conceived and the kind of information they use to compute the link weights. None of them investigated any weighting criterion that considered topological, temporal and contextual data simultaneously.

The work in [19] first investigated the use of link strength to the link prediction problem. The proposed weighting criterion was based exclusively on topological data: the frequency of existing interactions (i.e. the number of edges) between nodes. Formally, this criterion is defined as follows.

$$w(u, v) = |E(u, v)| \tag{1}$$

The work reported in [20] considers that link strength is already given by contextual attributes such as: number of flights between airports, weights between neurons in neural networks and number of common publications of two authors. Hence it does not present any initiative to combine different kinds of data for link strength computation. Although no weighting procedure is explicitly indicated, the work evaluates the effects of a parameter that intensifies/attenuates link strength.

Time-evolving link data can be modeled as a third order tensor. Based on this perspective, Dunlavy et al. [21] used a tensor decomposition method to extract unique three-dimensional components that represent data in sequential time slices. Then, for each component, the work applied the outer product of two vectors with contextual information in order to quantify the relationship (link strength) between object pairs. Although the proposed weighting criterion used contextual and temporal information to weight links, it did not consider topological data in link strength calculation.

In [16], the authors proposed three different weighting criteria: one for each kind of data (temporal, topological and contextual) observed in co-authorship networks. In the temporal-based criterion, the weight between two nodes was given by the year of the time-stamp of the most recent link between the nodes. The topological-based one was the frequency of existing interactions between two nodes (Eq. (1)). The contextual-based criterion considered the inverse of the minimum number of co-authors that collaborated in papers written by two nodes. The weighting criteria were evaluated independently and compared with each other. No one outperformed the others. The authors suggested criteria combination as future work.

Based on the principle of homophily, Xiang et al. [22] proposed a link-based latent variable unsupervised model to estimate relationship strength from user profile similarity (described by attributes like gender, marital status, political view, among others) and interaction activity such as communication and tagging. Although the work analyzed the weighted graph from different points of view, there was no direct comparison between the proposed model and any topology-based weighting criteria. Additionally, the temporal aspect was not taken into consideration in the weighting process.

The work in [18] investigated a weighting criterion that takes temporal information into account (see Eq. (2)). It returns the elapsed time (age) from the most recent interaction between two linked nodes $u$ and $v$ (given by $max(t_{(u,\ v)})$) to the current time ($CTime$). $\beta$ is an arbitrary damping factor ($0 < \beta \le 1$). Hence, weights between connected nodes that interacted recently are higher than the ones whose last interactions occurred before in the past. Moreover, the strength of a link varies over time. Links between two nodes that have not interacted with each other for a long time, with respect to the current time, become weaker. Muniz et al. [18] used this criterion to extend the topological information-based criterion described in Eq. (1). The extended criterion was evaluated and compared to the topology-based one. Results provided experimental evidence that the combination of topological and temporal information to weight links improves link prediction. Yet the proposed criterion did not take contextual information into account in the weighting process.

$$w(u, v) = \beta^{CTime - max(t_{(u,v)})} \tag{2}$$

Table 1 presents a comparative summary of the research in this scenario, indicating what kind of information was used for link strength calculation.

At last, it is important to emphasize the popularity of the frequency of existing interactions as a weighting criterion in the unsupervised link prediction area. We believe that such popularity is mainly due to its computational simplicity. Yet this criterion is limited to a single topological aspect of the networks. It does not consider other aspects observed in many social networks such as contextual and temporal information.

## 3. Proposed weighting criteria

Although several works on unsupervised link prediction used topological, temporal and contextual data to formulate weighting criteria, none of them combined these three aspects *simultaneously*. Hence, in this section, we initially propose a general weighting model that can be used to configure different weighting criteria based on combinations of contextual, temporal and topological information. Then, we configure the proposed model generating three specific weighting criteria, each one emphasizing different aspects observed in social networks.

**Table 1**
Weighting-based unsupervised link prediction–related work summary.

| | Kind ofdata used in link strength calculation | | |
|---|---|---|---|
| Reference | Topological | Temporal | Contextual |
| [19] | yes | no | no |
| [20] | no | no | yes |
| [21] | no | yes | yes |
| [16] | yes | yes | yes |
| [22] | yes | no | yes |
| [18] | yes | yes | no |

### *3.1. General model*

Consider a social network represented by a graph $G(V, E)$ as stated in Section 2.1. Suppose that $G$ may contain attributes (i.e. contextual data) in nodes as well as in edges. State-of-the-art in unsupervised link prediction offers a plethora of similarity functions that can be used for weight calculation in this scenario (Section 2.2). Consider three sets of those functions: *Top, Tem* and *Con*, organized according to the kind of information used in weight computation. These sets contain topological, temporal and contextual-based weighting criteria, respectively. Given two arbitrary criteria $d_i$, $d_j \in Top \cup Temp \cup Con$, a new weighting criteria $d$: $V \times V \to \mathbb{R}$ can be defined as $d(u, v) = d_i(u, v) \times d_j(u, v)$. Product between $d_i$ and $d_j$ ensures that both weighting criteria are considered simultaneously. Hence we propose our general weighting model as stated in Eq. (3), where $top \in Top$, $tem \in Tem$, $con \in Con$ and $x_{top}$, $x_{tem}$, $x_{con} \in \{0, 1\}$.

$$w^*(u, v) = top^{x_{top}} \times tem^{x_{tem}} \times con^{x_{con}} \tag{3}$$

The proposed model allows the configuration of different weighting criteria. This configuration has two levels. In the first, one can choose what kind of data must be taken into consideration in edge weighting: topological, temporal and contextual. It is defined by setting the desired flags ($x_{top}$, $x_{tem}$, $x_{con}$) to 1 and the others to 0. In the second level, similarity functions *top, tem* and *con* must be chosen. It is important to emphasize that the product between weighting criteria in model formulation ensures that the selected aspects ($x_i = 1$) must be considered simultaneously.

In face of the great diversity of similarity functions, the above-mentioned choice is certainly not an easy task to the analyst. She must be aware of the existing functions and the theories they are based on. In this sense, this choice depends mostly on her knowledge about the problem and on what aspects observed in the social network (topological/temporal/contextual) she wants to take into account.

In this paper, we are particularly interested in evaluating whether combining contextual, temporal and topological information in weight calculation can improve topological exclusive weighting-based link prediction. Therefore, the next sub-sections describe three weighting criteria configured from our proposed model: Temporal & Topological, Contextual & Topological and All Feature.

### *3.2. Temporal & topological weighting*

The temporal-topological weighting criterion (TT) is inspired by the Weak Ties theory, which states that recent and intense interactions tend to stimulate new interactions in the network [29]. The idea is to combine data about the time and the frequency of interactions in order to predict links. Therefore, recent and recurrent (intense) interactions must have higher influence than old and seldom ones in link prediction calculation. In order to implement TT according to the above-mentioned rationale, the general weighting model can be configured as indicated in Table 2.

Hence, the *TT* weighting criterion for an interacting node pair $u$ and $v$ is thus defined as follows:

$$w^{TT}(u, v) = |E(u, v)|*\beta^{CTime-max(t_{(u,v)})} \tag{4}$$

where $\beta$ is an arbitrary damping parameter ($0 < \beta \leq 1$) used to calibrate the importance of time in the weighting criterion. Higher (resp. lower) values of $\beta$ intensify (resp. attenuate) influence of time in weight definition.

**Table 2**
TT weighting criterion–parameter configuration.

| Parameter | Value | Comment |
|---|---|---|
| $x_{top}$ | 1 | — |
| $x_{tem}$ | 1 | — |
| $x_{con}$ | 0 | — |
| ***top*** | Frequency of existing interactions between nodes (Eq. (1)). | It represents the intensity of interactions between the nodes. |
| ***tem*** | Age of the most recent timestamp (Eq. (2)). | It distinguishes between nodes with recent interactions from nodes with past interplay. |
| ***con*** | — | — |

Consider the co-authorship network[2] example presented in Fig. 2. If we use: (a) $\beta = 0.8$, (b) CTime = 2017, and (c) the weighted similarity function *WCN* for score calculation, the weighting step of the unsupervised link prediction process described in Section 2.1 would output:

$$w^{TT}(Ava, Dana) = 3*0.8^{(2017-2017)} = 3*1 = 3.0$$

$$w^{TT}(Bob, Dana) = 3*0.8^{(2017-2016)} = 3*0.8 = 2.4$$

$$w^{TT}(Cal, Dana) = 3*0.8^{(2017-2015)} = 3*0.64 = 1.9$$

$$WCN^{TT}(Ava, Bob) = (3 + 2.4)/2 = 2.70$$

$$WCN^{TT}(Ava, Cal) = (3 + 1.9)/2 = 2.45$$

$$WCN^{TT}(Bob, Cal) = (2.4 + 1.9)/2 = 2.15$$

On the other hand, if the weighting criterion was exclusively based on topological information, there would be no prevalence for link prediction and the final score would be the same for all pairs of nodes:

$$w^T(Ava, Dana) = w^T(Bob, Dana) = w^T(Cal, Dana) = 3.0$$

$$WCN^T(Ava, Bob) = WCN^T(Ava, Cal) = WCN^T(Bob, Cal) = 3.0$$

However, using $w^{TT}$ (i.e. combining temporal and topological information provided in the network), the pair (*Ava, Bob*) would be more likely to connect than the others. Indeed, although the three authors have the same frequency of interaction with their common neighbor, *Ava* and *Bob* are the authors that most recently interacted with that neighbor. Hence, they should be more strongly linked to such neighbor than others. The *TT*'s temporal-based factor depicted this aspect. Consequently, (*Ava, Bob*) received the highest weighted common neighbor score among all pairs of non-connected nodes.

### *3.3. Contextual & topological weighting*

The CT weighting criterion is inspired by the emergence of the homophily theory [31] in social networks. This social theory states that individuals that share interests and/or present similar characteristics (i.e. have common contextual information) tend to associate with each other. Hence, the idea of this criterion is to combine similarity between nodes (homophily) with the frequency (intensity) of their interactions so that connected nodes that interact frequently and share similar contextual information have higher link strength. Table 3 presents *CT*'s parameter setting.

The *CT* weighting criterion for a node pair $u$ and $v$ is thus defined as follows:

$$w^{CT}(u, v) = |E(u, v)|*\alpha^{(1-cos(u,v))} \tag{5}$$

where $\alpha$ is an arbitrary damping parameter ($0 < \alpha \leq 1$) used to calibrate the importance of contextual information in the weighting criterion. Higher (resp. lower) values of $\alpha$ intensify (resp. attenuate) influence of contextual information in weight definition.

Consider again the example from Fig. 2 and that $\alpha = 0.5$ and $CTime = 2017$. Suppose that, in this case, the aggregation function $f$ for

[2] Co-authorship networks are highly dynamic social networks since several papers are published every year. The frequently upcoming publications are normally associated with new authors and/or new collaborations [5]. The link prediction problem, in such networks, consists of deducing if two authors will establish a coauthor relationship in the near future.

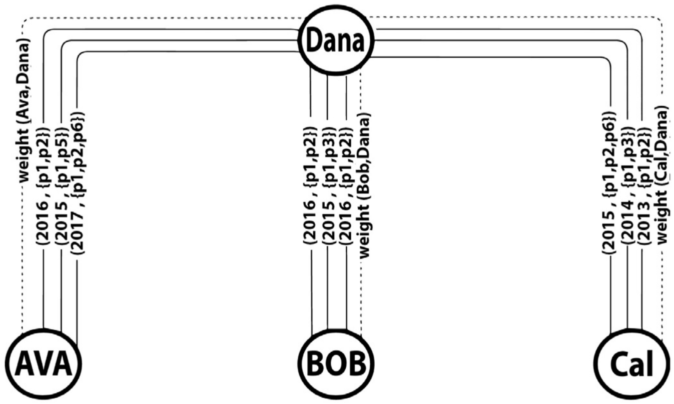


**Fig. 2.** A sample co-authorship network represented as an augmented graph with both structure and attribute information: nodes and continuous edges denote authors and papers (i.e. collaborations), respectively. Each paper contains two attributes: year of publication (temporal) and set of keywords (contextual). Each dashed line is an artificial edge created by the weighting step and contains the link strength between the corresponding nodes.

**Table 3**
CT weighting criterion–parameter configuration.

| Parameter | Value | Comment |
|---|---|---|
| $x_{top}$ | 1 | — |
| $x_{tem}$ | 0 | — |
| $x_{con}$ | 1 | — |
| ***top*** | Frequency of existing interactions between nodes (Eq. (1)). | It represents the intensity of interactions between the nodes. |
| ***tem*** | — | — |
| ***con*** | Cosine similarity (Section 2.1) | It expresses how much two nodes share contextual information. |

the cosine similarity (see Section 3.1) is the union of the keywords from all papers of a node. Then, we would have:

$$f\,(Ava) = \{p_1,\, p_2\} \cup \{p_1,\, p_5\} \cup \{p_1,\, p_2,\, p_6\} = \{p_1,\, p_2,\, p_5,\, p_6\}$$

$$f\,(Bob) = \{p_1,\, p_2\} \cup \{p_1,\, p_2\} \cup \{p_1,\, p_3\} = \{p_1,\, p_2,\, p_3\}$$

$$f\,(Cal) = \{p_1,\, p_2\} \cup \{p_1,\, p_3\} \cup \{p_1,\, p_2,\, p_6\} = \{p_1,\, p_2,\, p_3,\, p_6\}$$

$$\begin{aligned} f\,(Dana) \;&= \{p_1,\, p_2\} \cup \{p_1,\, p_5\} \cup \{p_1,\, p_2,\, p_6\} \cup \{p_1,\, p_2\} \cup \{p_1,\, p_3\} \cup \\ &\{p_1,\, p_2\} \cup \{p_1,\, p_3\} \cup \{p_1,\, p_2,\, p_6\} = \{p_1,\, p_2,\, p_3,\, p_5,\, p_6\} \end{aligned}$$

$$cos\,(Ava,\, Dana) = \frac{|\{p_1,\, p_2,\, p_5,\, p_6\}|}{\sqrt{|\{p_1,\, p_2,\, p_5,\, p_6\}|} \times \sqrt{|\{p_1,\, p_2,\, p_3,\, p_5,\, p_6\}|}} = 0.89$$

$$cos\,(Bob,\, Dana) = \frac{|\{p_1,\, p_2,\, p_3\}|}{\sqrt{|\{p_1,\, p_2,\, p_3\}|} \times \sqrt{|\{p_1,\, p_2,\, p_3,\, p_5,\, p_6\}|}} = 0.77$$

$$cos\,(Cal,\, Dana) = \frac{|\{p_1,\, p_2,\, p_3,\, p_6\}|}{\sqrt{|\{p_1,\, p_2,\, p_3,\, p_6\}|} \times \sqrt{|\{p_1,\, p_2,\, p_3,\, p_5,\, p_6\}|}} = 0.89$$

$$w^{CT}\,(Ava,\, Dana) = 3{*}0.93 = 2.79$$

$$w^{CT}\,(Bob,\, Dana) = 3{*}0.86 = 2.57$$

$$w^{CT}\,(Cal,\, Dana) = 3{*}0.93 = 2.79$$

$$WCN^{CT}\,(Ava,\, Bob) = (2.79 + 2.57)/2 = 2.68$$

$$WCN^{CT}\,(Ava,\, Cal) = (2.79 + 2.79)/2 = 2.79$$

$$WCN^{CT}\,(Bob,\, Cal) = (2.57 + 2.79)/2 = 2.68$$

As previously observed, using the traditional frequency of existing interactions weight, there would be no prevalence for link prediction. Nevertheless, using $w^{CT}$ (i.e. using contextual and topological information provided in the network), the pair (*Ava, Cal*) would be more likely to connect than the others. Indeed, although the three authors have the same frequency of interaction with their common neighbor, *Ava* and *Cal* are the authors that share the highest number of common keywords with their common neighbor (i.e. they have similar interests in research as their common neighbor). Hence, they should be more strongly connected to such neighbor than others. *CT*'s contextual-based factor depicted this aspect. Consequently, (*Ava, Cal*) received the highest weighted common neighbor score among all pairs of non-connected nodes.

### 3.4. All feature weighting

The third weighting criterion (CTT) considers contextual, temporal and topological aspects of the network simultaneously. It merges TT and CT criteria. Hence, CTT takes into account principles from the Weak Ties and the Homophily social theories. The idea is to combine profile similarity between nodes (homophily) with frequency (intensity) and time (age) of their interactions so that connected nodes that interacted frequent and recently and share similar profiles (i.e., present common interests) have higher link strength. Table 4 contains the general weighting model's configuration w.r.t. CTT.

The all-feature weighting score for a node pair $u$ and $v$ is thus defined as follows:

$$w^{CTT}\,(u,\, v) = |E\,(u,\, v)|{*}\beta^{\frac{CTime - max\left(t_{(u,v)}\right)}{CTime - min\,(t)}}{*}\alpha^{1 - cos\,(u,v)} \tag{6}$$

where $\alpha$ and $\beta$ are arbitrary damping factors ($0 < \alpha,\ \beta \le 1$) to put more/less emphasis on the contextual and temporal influences respectively. For scaling purposes, we normalize the age of the most recent time-stamp with the oldest age ($Ctime\, -\, min\,(t)$) among all interactions.

Considering once again the example from Fig. 2 and using: (a) $\beta = 0.5$; (b) $\alpha = 0.5$; (c) current time $=$ 2017; (d) aggregation function $f$ $=$ union of keywords, we would have:

$$w^{CTT}\,(Ava,\, Dana) = 3{*}0.\ 8^{\frac{(2017 - max\,(2017,2016,2015))}{2017 - 2013}}{*}0.\ 5^{1 - 0.89} \approx 2.79$$

$$w^{CTT}\,(Bob,\, Dana) = 3{*}0.\ 8^{\frac{(2017 - max\,(2016,2015,2016))}{2017 - 2013}}{*}0.\ 5^{1 - 0.77} \approx 2.16$$

**Table 4**
CTT weighting criterion–parameter configuration.

| Parameter | Value | Comment |
|---|---|---|
| $x_{top}$ | 1 | — |
| $x_{tem}$ | 1 | — |
| $x_{con}$ | 1 | — |
| ***top*** | Frequency of existing interactions between nodes (Eq. (1)). | It represents the intensity of interactions between the nodes. |
| ***tem*** | Age of the most recent time-stamp (Eq. (2)). | It distinguishes between nodes with recent interactions from nodes with past interplay. |
| ***con*** | Cosine similarity (Section 2.1) | It expresses how much two nodes share contextual information. |

$$w^{CTT}(Cal, Dana) = 3{*}0.\ 8^{\frac{(2017-max(2017,2015,2016))}{2017-2013}}{*}0.\ 5^{1-0.89} \approx 1,\ 97$$

$$WCN^{CTT}(Ava, Bob) \approx (2.79 + 2.16)/2 \approx 2.47$$

$$WCN^{CTT}(Ava, Cal) \approx (2.79 + 1.97)/2 \approx 2.38$$

$$WCN^{CTT}(Bob, Cal) \approx (2.16 + 1.97)/2 \approx 2.06$$

Under the temporal-topological point of view, (Ava, Bob) should be ranked first (see Section 3.2). Considering the contextual-topological aspect, (Ava, Cal) should win (see Section 3.2). *CTT* considers both points of view simultaneously. In the example, it was configured to emphasize them equally ($\beta = 0.5$ and $\alpha = 0.5$). According to this configuration, (Ava, Bob) would reach the top-1 position for link recommendation. Indeed, Ava and Bob are the ones that better satisfy the above-mentioned points of view at the same time: they most frequent and recently interacted with their common neighbor and also shared common keywords with him.

## 4. Experiments

### 4.1. Datasets

We used two versions of five co-authorship networks, in our experiments. Such networks were used in [2] and describe the authors and papers from five sections of the physics e-Print arXiv: astro-ph (astrophysics), cond-mat (condensed matter), gr-qc (general relativity and quantum cosmology), hep-ph (high energy physics–phenomenology) and hep-th (high energy physics–theory). The first version covered the same time interval (papers from 1994 to 1999) used by Liben-Nowell and Kleinberg [2]. That was very important to help us validate our implementation. The second version covered the same period (papers from 2000 to 2005) used by Munasinghe and Ichise [23]. All networks were extracted from arXiv API.[3] Both versions of the networks were homogeneous attributed multigraphs where nodes and edges represent authors and papers, respectively. All networks contained two attributes in edges: the paper's year of publication and its set of keywords.

### 4.2. Experimental methodology

Our experiments employed the weighting-based procedure depicted in Fig. 1. First, we divided each network in two periods of three years. Hence, networks from the first set were partitioned into $G_{Trn} = [1994, 1996]$ and $G_{Tst} = [1997, 1999]$ and networks from the second set were split into $G_{Trn} = [2000, 2002]$ and $G_{Tst} = [2003, 2005]$.

Then, we proceeded to perform the experiment configuration. In this configuration, the user can choose the experiment settings and fine-tune the parameters of the weighting criteria. For score calculation, *WCN* and *WAA* were chosen because they are the similarity functions most used in weighting-based link prediction studies. Each of them was combined with the following weighting criteria: *T* (exclusively topological–the frequency of existing interactions), *TT* (Temporal & Topological), *CT* (Contextual & Topological) and *CTT* (All Feature). Hence, we considered eight similarity functions in our experiments: $WCN^T$, $WAA^T$, $WCN^{TT}$, $WAA^{TT}$, $WCN^{CT}$, $WAA^{CT}$, $WCN^{CTT}$ and $WAA^{CTT}$. It is important to emphasize that we used the union of keywords as the aggregation function *f* for the cosine similarity in all cases. To calibrate parameters $\alpha$ and $\beta$, we divided $G_{Trn}$ of each network from the first group into two subsets: $G_{Pre} = [1994, 1995]$ and $G_{Validation} = [1996, 1996]$ and $k = 1$. We ranged the values of $\alpha$ and $\beta$ from 0.1 to 0.9 with step 0.1. We picked up the values that led to the best results in the validation set (see Table 5 for the complete configuration adopted in the experiment).

**Table 5**
Experimental configuration set.

| Similarity function | Parameters | |
|---|---|---|
| | $\alpha$ | $\beta$ |
| $WCN^T$ | – | – |
| $WAA^T$ | – | – |
| $WCN^{TT}$ | – | 0.2 |
| $WAA^{TT}$ | – | 0.4 |
| $WCN^{CT}$ | 0.5 | – |
| $WAA^{CT}$ | 0.5 | – |
| $WCN^{CTT}$ | 0.7 | 0.2 |
| $WAA^{CTT}$ | 0.7 | 0.4 |

To identify the nodes that belong to the `Core` set (see process for weighting-based link prediction, activity 2, described in Section 2.1), we considered $k = 3$. Hence, `Core` consisted of all active authors who had written at least 3 articles during the training period and at least 3 articles during the test period. Three reasons guided this choice: (a) Training and test periods' length of all networks was three years; (b) We considered that one year could be a reasonable frequency interval for paper publication; (c) It was the same value defined in [2], where similar experiments were performed.

For the graph weighting step, we created artificial edges between nodes connected in $G_{Trn}$. Then we calculated four weight values (one for each weighting criterion) for each artificial edge.

Finally, we executed each similarity function for each pair of non-connected nodes in each network and compared the performances of all similarity functions to the performance of the random predictor (see improvement factor calculation described in Section 2.1).

### 4.3. Results

Tables 6 and 7 provide some statistics of the networks after the Core Identification step.[4] Tables 8 and 9 show the functions' performances on each network with respect to the improvement factor over the random predictor ($p_{rand}$). Best values are highlighted in bold font. A preliminary analysis reveals that no function outperformed all the others in all networks and periods. Nevertheless, it is important to emphasize that *CTT* won in 6 out of 10 networks. If we consider criteria with any kind of combination, the number of victories grows to 8 cases and 2 ties. The traditional topology-based criterion *T* never won alone. When it won, its performances were also achieved by a combination-based criteria. Hence, in general, the results suggest that combining topological, temporal and contextual data may, in fact, be an interesting choice to enhance link prediction.

On the other hand, a more rigorous analysis reveals that, in most cases, the differences of performance among the similarity functions were not high. Hence, in order to check for statistical significance in those differences, we applied the Friedman test with significance level $\alpha = 0.05$ [32] and null hypothesis stated as $H_0$: "the performance of the similarity functions are statistically identical". In the test, we considered the eight similarity functions and the ten networks. The test rejected $H_0$, indicating a significant statistical difference among the functions.

In order to better investigate the differences in performance of the similarity functions, we decided to run a post-hoc test to compare the functions with each other. The Nemenyi test with significance level $\alpha = 0.05$ [32] was employed. We stated $H_0$ as "performances of functions $x$ and $y$ are statistically identical". We run the test twenty eight times, one for each possible pair of functions. Ten out of the twenty eight tests revealed statistical difference, i.e. $H_0$ was rejected. Table 10

[3] http://export.arxiv.org/api/ .

[4] As previously presented in Section 2.1, `Core` is the set of active nodes and $E_{New}$ is the set of edges to be foreseen.

**Table 6**
Statistics about the 1st version of the networks used in the experiments (1994–1999).

| Network | Authors | Papers | Core | $E_{New}$ |
|---|---|---|---|---|
| astro-ph | 19,864 | 21,290 | 9616 | 2087 |
| cond-mat | 19,289 | 21,698 | 1336 | 723 |
| **gr-qc** | 5283 | 8299 | 390 | 137 |
| hep-ph | 12,658 | 24,294 | 1689 | 1950 |
| hep-th | 11,229 | 20,935 | 1192 | 767 |

**Table 7**
Statistics about the 2nd version of the networks used in the experiments (2000–2005).

| Network | Authors | Papers | Core | $E_{New}$ |
|---|---|---|---|---|
| astro-ph | 42,771 | 50,359 | 6197 | 37,362 |
| cond-mat | 48,298 | 51,809 | 4437 | 7507 |
| gr-qc | 8939 | 13,858 | 812 | 463 |
| hep-ph | 17,750 | 31,707 | 2476 | 8246 |
| **hep-th** | 14,212 | 27,444 | 1893 | 1293 |

**Table 8**
Improvement factor of similarity function over the random predictor–1st version of the networks (1994–1999).

| | Network | | | | |
|---|---|---|---|---|---|
| Similarity function | astro-ph | cond-mat | gr-qc | hep-ph | hep-th |
| $Rand_{Pred}$ | 0.23 | 0.11 | 0.18 | 0.14 | 0.11 |
| $WCN^{T}$ | 27.9 | 64.9 | **68.5** | 44.8 | **93.9** |
| $WAA^{T}$ | 40.0 | 69.9 | 64.4 | 47.5 | **93.9** |
| $WCN^{TT}$ | 34.5 | 59.9 | 64.4 | 45.6 | 77.1 |
| $WAA^{TT}$ | 40.2 | 67.4 | **68.5** | 56.4 | 77.1 |
| $WCN^{CT}$ | 39.8 | 64.9 | 40.3 | 48.6 | 61.4 |
| $WAA^{CT}$ | 36.3 | 68.6 | 56.4 | **52.7** | 60.2 |
| $WCN^{CTT}$ | 38.7 | 71.1 | 64.4 | 49.7 | 75.9 |
| $WAA^{CTT}$ | **41.0** | **76.1** | 64.4 | 50.5 | 78.3 |

**Table 9**
Improvement factor of similarity function over the random predictor–2nd version of the networks (2000–2005).

| | Network | | | | |
|---|---|---|---|---|---|
| Similarity function | astro-ph | cond-mat | gr-qc | hep-ph | hep-th |
| $Rand_{Pred}$ | 0.20 | 0.08 | 0.14 | 0.27 | 0.07 |
| $WCN^{T}$ | 50.2 | 106.2 | 45.9 | 61.2 | 83.4 |
| $WAA^{T}$ | 52.5 | 116.3 | 50.5 | 61.7 | 83.4 |
| $WCN^{TT}$ | 51.2 | 112.0 | 49.0 | 62.3 | 78.1 |
| $WAA^{TT}$ | **54.1** | 118.8 | 52.1 | 62.6 | 86.6 |
| $WCN^{CT}$ | 51.5 | 99.6 | 33.7 | 52.8 | 66.3 |
| $WAA^{CT}$ | 52.0 | 107.1 | 39.8 | 60.2 | 72.7 |
| $WCN^{CTT}$ | 53.5 | 116.3 | 50.5 | **64.0** | 86.6 |
| $WAA^{CTT}$ | 53.5 | **120.9** | **53.6** | 63.7 | **94.1** |

summarizes them. For each test, the table indicates which function of the corresponding pair showed higher performance.

In 90% of the tests that presented significant differences, similarity functions based on combined weighting criteria outperformed the others. These results confirm our hypothesis that combining contextual, temporal and topological information can enhance unsupervised link prediction when it is based exclusively on topological data.

The most prevalent combined weighting criteria in tests with statistical differences were *CTT* (tests 6–10) and *TT* (tests 2–5). Hence, the combination of temporal and topological information was present in 90% of all tests presented in Table 10. On the other hand, *CT* was the only combined weighting criterion that did not outperform any other significantly. Moreover, it was the only criterion outperformed by a purely topological weighting criterion (test 1). Despite there is not a plausible explanation for *CT*'s poor performance, if we consider all the results, it seems that temporal information may have played a key role to determine the outcomes of this experiment (it won 55% of the tests with statistical differences). This hypothesis is reinforced by two facts. First, although *CT* presented a low performance, *CTT* was the criterion with the best results. Second, *CTT* is an extended version of *CT* where temporal information is added to the combination of contextual and topological data.

**Table 10**
Pairs of similarity functions that revealed a statistically significant difference in performance (according to the Nemenyi Test)–each row indicates which function of the corresponding pair showed higher performance.

| | Similarity function performance | |
|---|---|---|
| Test | Higher | Lower |
| 1 | $WAA^{T}$ | $WCN^{CT}$ |
| 2 | $WAA^{TT}$ | $WCN^{T}$ |
| 3 | $WAA^{TT}$ | $WCN^{CT}$ |
| 4 | $WAA^{TT}$ | $WAA^{CT}$ |
| 5 | $WAA^{TT}$ | $WCN^{TT}$ |
| 6 | $WAA^{CTT}$ | $WCN^{T}$ |
| 7 | $WAA^{CTT}$ | $WCN^{CT}$ |
| 8 | $WAA^{CTT}$ | $WAA^{CT}$ |
| 9 | $WAA^{CTT}$ | $WCN^{TT}$ |
| 10 | $WCN^{CTT}$ | $WCN^{CT}$ |

Another important aspect to be emphasized is that the Weighted Adamic Adar index (*WAA*) won 90% of the tests. It seems that the participants of the co-authorship networks employed in the experiment were restrictive when choosing their collaborators.

## 5. Conclusion and future work

Predicting whether a pair of nodes will connect in the future is an important network analysis task known as the link prediction problem. One of the major approaches to the this problem computes link weights between connected nodes and, based on a weighted graph, apply weighted similarity functions between non-connected nodes in order to identify potential new links. The weighting criteria commonly adopted by related studies were based exclusively on topological information, i.e. information that describes structural aspects of the network analyzed. Nevertheless, such approach leads to poor incorporation of other aspects of the social networks, such as context (node and link attributes), and temporal information (chronological interaction data). Hence, in this article, we investigated whether the combination of contextual and temporal information with topological data in weight computation could improve the performance of link prediction methods. Our proposal includes a general weighting model that allows the user to configure different weighting criteria based on combinations of contextual, temporal and topological information. It also includes three graph weighting criteria configured from the general model in order to implement such combinations:

- The Temporal-Topological (*TT*) criterion combines the frequency of interactions (topological) between connected nodes and the age of the most recent interaction (temporal).
- The Contextual-Topological criterion (*CT*) merges the similarity between the profiles (contextual) of connected nodes and the frequency of interactions (topological) between those nodes.
- The all feature criterion (*CTT*) gathers frequency (topological) and age of interactions (temporal) with similarity between the profiles (contextual) of connected nodes.

*CTT* was the article's main contribution. It combines topological, temporal and contextual information *simultaneously*. Experimental results with two popular weighted similarity functions (Adamic-Adar Index and Common Neighbors) in ten social networks provided statistical evidence that *CTT* does enhance unsupervised link prediction when compared to other weighting criteria that do not combine the three aspects. It confirmed our hypothesis that combining topological, contextual and temporal aspects of social networks in weight calculation can, indeed, enhance unsupervised link prediction.

As future work, we consider evaluating other weighting criteria configured from our general weighting model. We also plan to develop an optimization procedure to support weighting criteria parameter configuration. Additionally, it would be interesting to evaluate the influence of our criteria in the supervised approach to the link prediction problem. Experiments of our criteria with networks out of the context of co-authorship would be desirable too.

## References

[1] R. Zeng, Y.X. Ding, X.L. Xia, Link prediction based on dynamic weighted social attribute network, Proceedings of the 2016 International Conference on Machine Learning and Cybernetics, volume 1 of ICMLC, (2016), pp. 183–188, http://dx.doi.org/10.1109/ICMLC.2016.7860898.

[2] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Am. Soc. Inf. Sci. Technol. 58 (7) (2007) 1019–1031, http://dx.doi.org/10.1002/asi.v58:7.

[3] L. Zhu, D. Guo, J. Yin, G.V. Steeg, A. Galstyan, Scalable temporal latent space inference for link prediction in dynamic social networks, IEEE Trans. Knowl. Data Eng. 28 (10) (2016) 2765–2777, http://dx.doi.org/10.1109/TKDE.2016.2591009.

[4] R.-H. Li, J.X. Yu, J. Liu, Link prediction: the power of maximal entropy random walk, Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11, (2011), pp. 1147–1156, http://dx.doi.org/10.1145/2063576.2063741.

[5] C. Zhang, H. Zhang, D. Yuan, M. Zhang, Deep learning based link prediction with social pattern and external attribute knowledge in bibliographic networks, Proceedings of the 2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), (2016), pp. 815–821, http://dx.doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2016.170.

[6] L. Lü, T. Zhou, Link prediction in complex networks: a survey, Physica A 390 (6) (2011) 1150–1170, http://dx.doi.org/10.1016/j.physa.2010.11.027.

[7] M.A. Hasan, M.J. Zaki, A survey of link prediction in social networks, in: C.C. Aggarwal (Ed.), Social Network Data Analytics, Springer US, Boston, MA, 2011, pp. 243–275, , http://dx.doi.org/10.1007/978-1-4419-8462-3_9.

[8] V. Martínez, F. Berzal, J.-C. Cubero, A survey of link prediction in complex networks, ACM Comput. Surv. 49 (4) (2016) 69:1–69:33, http://dx.doi.org/10.1145/3012704.

[9] Z.L. Li, X. Fang, O.R.L. Sheng, A survey of link recommendation for social networks: methods, theoretical foundations, and future research directions, ACM Trans. Manage. Inf. Syst. 9 (1) (2017) 1:1–1:26, http://dx.doi.org/10.1145/3131782.

[10] P. Wang, B. Xu, Y. Wu, X. Zhou, Link prediction in social networks: the state-of-the-art, Sci. China Inf. Sci. 58 (1) (2015) 1–38, http://dx.doi.org/10.1007/s11432-014-5237-y.

[11] M.A. Hasan, V. Chaoji, S. Salem, M. Zaki, Link prediction using supervised learning, Proceedings of the SDM 06 workshop on Link Analysis, Counterterrorism and Security, (2006), pp. 1–10.

[12] P.R.S. Soares, R.B.C. PrudêNcio, Proximity measures for link prediction based on temporal events, Expert Syst. Appl. 40 (16) (2013) 6652–6660, http://dx.doi.org/10.1016/j.eswa.2013.06.016.

[13] C.A. Bliss, M.R. Frank, C.M. Danforth, P.S. Dodds, An evolutionary algorithm approach to link prediction in dynamic social networks, J. Comput. Sci. 5 (5) (2014) 750–764, http://dx.doi.org/10.1016/j.jocs.2014.01.003.

[14] P. Bhattacharyya, A. Garg, S.F. Wu, Analysis of user keyword similarity in online social networks, Soc. Netw. Anal. Min. 1 (3) (2011) 143–158, http://dx.doi.org/10.1007/s13278-010-0006-4.

[15] L.M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, F. Menczer, Friendship prediction and homophily in social media, ACM Trans. Web 6 (2) (2012) 9:1–9:33, http://dx.doi.org/10.1145/2180861.2180866.

[16] T. Tylenda, R. Angelova, S. Bedathur, Towards time-aware link prediction in evolving social networks, Proceedings of the 3rd Workshop on Social Network Mining and Analysis, SNA-KDD'09, (2009), pp. 1–10, http://dx.doi.org/10.1145/1731011.1731020.

[17] P. Choudhary, N. Mishra, S. Sharma, R. Patel, Link score: a novel method for time aware link prediction in social network, in: D. Prasad, D. Nalini (Eds.), Emerging Research in Computing, Information, Communication and Applications, Elsevier, 2013, pp. 111–118.

[18] C.P. Muniz, R. Choren, R. Goldschmidt, Using a time based relationship weighting criterion to improve link prediction in social networks, Proceedings of the 19th International Conference on Enterprise Information Systems, ICEIS'17, (2017), pp. 73–79.

[19] T. Murata, S. Moriyasu, Link prediction of social networks based on weighted proximity measures, Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, (2007), pp. 85–88, http://dx.doi.org/10.1109/WI.2007.52.

[20] L. Lü, T. Zhou, Link prediction in weighted networks: the role of weak ties, EPL (Europhys. Lett.) 89 (1) (2010) 18001, http://dx.doi.org/10.1209/0295-5075/89/18001.

[21] D.M. Dunlavy, T.G. Kolda, E. Acar, Temporal link prediction using matrix and tensor factorizations, ACM Trans. Knowl. Discov. Data 5 (2) (2011), http://dx.doi.org/10.1145/1921632.1921636.

[22] R. Xiang, J. Neville, M. Rogati, Modeling relationship strength in online social networks, Proceedings of the 19th International Conference on World Wide Web, WWW '10, ACM, New York, NY, USA, 2010, pp. 981–990, http://dx.doi.org/10.1145/1772690.1772790.

[23] L. Munasinghe, R. Ichise, Time score: a new feature for link prediction in social networks, IEICE Trans. Inf. Syst. E95.D (3) (2012) 821–828, http://dx.doi.org/10.1587/transinf.E95.D.821.

[24] A. Pecli, M.C.R. Cavalcanti, R.R. Goldschmidt, Automatic feature selection for supervised learning in link prediction applications: a comparative study, Knowl. Inf. Syst. (2017) 1–37.

[25] S. Scellato, A. Noulas, C. Mascolo, Exploiting place features in link prediction on location-based social networks, Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, ACM, New York, NY, USA, 2011, pp. 1046–1054, http://dx.doi.org/10.1145/2020408.2020575.

[26] S.-H. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng, H. Zha, Like like alike: joint friendship and interest propagation in social networks, Proceedings of the 20th International Conference on World Wide Web, WWW '11, ACM, New York, NY, USA, 2011, pp. 537–546, http://dx.doi.org/10.1145/1963405.1963481.

[27] M. Rowe, M. Stankovic, H. Alani, Who will follow whom? exploiting semantics for link prediction in attention-information networks, in: P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J.X. Parreira, J. Hendler, G. Schreiber, A. Bernstein, E. Blomqvist (Eds.), The Semantic Web – ISWC 2012, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 476–491.

[28] J. Tang, M. Musolesi, C. Mascolo, V. Latora, Temporal distance metrics for social network analysis, Proceedings of the 2Nd ACM Workshop on Online Social Networks, WOSN '09, (2009), pp. 31–36, http://dx.doi.org/10.1145/1592665.1592674.

[29] M.S. Granovetter, The strength of weak ties, Am. J. Sociol. 78 (6) (1973) 1360–1380, http://dx.doi.org/10.1086/225469.

[30] J. Valverde-Rebaza, A. Valejo, L. Berton, T. de Paulo Faleiros, A. de Andrade Lopes, A Naïve Bayes model based on overlapping groups for link prediction in online social networks, Proceedings of the 30th Annual ACM Symposium on Applied Computing, ACM SAC '15, (2015), pp. 1136–1141, http://dx.doi.org/10.1145/2695664.2695719.

[31] S. Currarini, M.O. Jackson, P. Pin, An economic model of friendship: homophily, minorities, and segregation, Econometrica 77 (4) (2009) 1003–1045, http://dx.doi.org/10.3982/ECTA7528.

[32] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.