Link Prediction in Evolutionary Graphs

The Case Study of the CCIA Network

Kemo ADRIAN [a], Paula CHOCRON [a], Roberto CONFALONIERI [a], Xavier FERRER [a] and Jesús GIRÁLDEZ-CRU [b]

[a] *Artificial Intelligence Research Institute, IIIA-CSIC Bellaterra (Barcelona), Catalonia, Spain*
{kemo.adrian,pchocron,confalonieri,xferrer}@iiia.csic.es

[b] *KTH Royal Institute of Technology, Stockholm, Sweden*
giraldez@kth.se

Abstract. Studying the prediction of new links in evolutionary networks is a captivating question that has received the interest of different disciplines. Link prediction allows to extract missing information and evaluate network dynamics. Some algorithms that tackle this problem with good performances are based on the sociability index, a measure of node interactions over time. In this paper, we present a case study of this predictor in the evolutionary graph that represents the CCIA co-authorship network from 2005 to 2015. Moreover, we present a generalized version of this sociability index, that takes into account the time in which such interactions occur. We show that this new index outperforms existing predictors. Finally, we use it in order to predict new co-authorships for CCIA 2016.

Keywords. Link prediction, Evolutionary networks, CCIA coauthorship network

1. Introduction

Graph models offer a natural framework to represent and analyze the interactions among the actors of complex real systems. As such, this modeling approach has been used in several real-world disciplines, including sociology, physics or biology. Many different tools and techniques to extract information from graphs have been developed in the past years, in an effort to respond to the needs of these applications. Most of these methods focus on the analysis of these graphs from a *static* point of view [6]. However, real-world systems are naturally dynamic and evolve in time, modifying their structure, e.g., by adding new edges. In such evolutionary graphs, the prediction of new links remains an interesting question, since these graphs are used to analyze relationships in the domains.

For these reasons, it is reasonable to tackle the problem of link prediction by analyzing the clustering evolution, i.e., how actors in an evolutionary graph are organized into *clusters* over time. Intuitively, a cluster (or *community*) is a set of nodes that have a high number of interactions between them with respect to the rest of nodes in the network. A common approach for the analysis of clustering evolution consists in splitting the evolutionary graph into *timestamps* or snapshots, i.e., static graphs at specific moments, computing the clusters of each timestamp, and, finally, studying the changes between one timestamp and the next one(s).

An example of this approach is the one by Asur et al. [2], where they propose the *Sociability Index* (SoI) to predict new links in evolutionary networks. This index measures the ability of a node to interact with nodes of other clusters or communities. This value is computed for each node, and is determined based on the clusters to which it belongs over time. In that paper, the sociability index is used to predict future interactions between existing nodes in the network. In particular, the authors claim that the sociability indexes of two unconnected nodes are directly related to the probability that they will be connected in the future.

To test the previous hypothesis, they analyze the evolutionary co-authorship network of 28 top AI conferences between 1997 and 2006. In the corresponding graph, each author is represented as a node, each timestamp represents a year, and two authors are connected in one timestamp if they are co-authors of a paper that year. The results obtained show that this index behaves well in such a large network, improving the random predictor 250 times.

The contribution of this paper is two-fold. On the one hand, we consider the question whether the sociability index approach also works well in a small network.[1] To answer this question, we analyze the co-authorship network of the CCIA conferences between 2005 and 2015, i.e., the evolutionary graph of co-authors containing papers published in these editions of CCIA. We observe that the size of the network has an important impact on the performance of the SoI index.

On the other hand, we propose a new predictor based on a generalized version of the sociability index, which takes into account the time in which interactions occurred. Intuitively, recent interactions among nodes should have more relevance than older ones. For instance, an author that has been socially active in the last years should have a higher sociability index than an author that was social a long time ago. Our generalization of the sociability index, called *Weighted Sociability Index* (SoI_α), captures this behavior. This cannot be done with the original SoI.

In an exhaustive experimental evaluation, we analyze the performance of some variants of the SoI_α, which differ in the time weight function α they use, in the CCIA co-authorship network. Our experiments show that, in general, these variants outperform the predictions of links computed by the SoI in our case study.

Finally, we use the most accurate SoI_α predictor found in this paper to obtain predictions about the current edition of the CCIA 2016.

[1] A *small* network has a small number of nodes and/or edges (w.r.t. another network). In our case-study, the CCIA co-authorship network is small w.r.t. the top AI conferences co-authorship network.

2. Preliminaries

A graph $G = (V, E)$ consists of a set of vertexes V and a set of edges $E \subseteq V \times V$ connecting pairs of vertexes. For simplicity, we only consider this definition of undirected and unweighted graphs; extensions to directed and/or weighted graphs are direct.

An evolutionary graph $G^T = (\{V^i\}, \{E^i\})$, with $1 \leq i \leq T$, is a set of graphs, where V^i and E^i represent the sets of vertexes and edges at timestamp i.

A clustering method is an algorithm that computes a partition $P = \{P_1, \ldots, P_k\}$ over the vertexes $v \in V$ of a graph. In this work, we use the Louvain method [3] to compute this partition. The Louvain method is a modularity optimization algorithm. Modularity [10] measures how dense the edges are within clusters. Modularity is one of the most popular metrics in clustering methods. However, the analysis presented in this paper is relatively independent of the clustering method used (as in [2]). Therefore, other clustering methods (such as K^* [4] or MCL [5]) would also apply. In an evolutionary graph, a clustering method computes T partitions, one for each timestamp. In these graphs, we denote as $c_t(x)$ the cluster to which a node x belongs at timestamp t.

Dataset. In order to build the CCIA co-authorship network, we used the papers published in the CCIA editions from 2005 to 2015. We built an evolutionary graph as follows. Each edition of the conference represents a timestamp. Each author is represented as a node, and there is an edge between two nodes if they both were co-authors of one of those papers. Nodes and edges are assigned to each timestamp accordingly. For each timestamp, we computed its clustering partition. Clusters may appear or disappear in each new edition. Clusters identify groups of nodes that are highly connected. If a group of nodes does not have major changes in its behaviour between two timestamps, its nodes must be characterized by the same cluster. To ensure this property, we use the Jaccard's distance between timestamps.

The evolutionary graph of the CCIA co-authorship network contains 11 timestamps, 714 nodes and 1991 edges among all timestamps, and an average of 107.64 nodes and 181 edges per timestamp.

3. The Sociability Index on a Small Network

In this section, we evaluate the performance of a predictor based on the sociability index in a small network, namely the CCIA co-authorship graph. Our predictor is based on the sociability defined by Asur et al. [2]. To this end, we first summarize their analysis.

3.1. Asur et al.'s Approach

Asur et al. [2] make predictions over a co-authorship network of 28 top AI international conferences in the period 1997-2006 with a total of 23136 nodes (authors) and 54989 edges.[2] In their analysis, first, they split the set of timestamps into

[2]Let us remark that each of those 28 conferences has a size greater than the CCIA (in term of number of accepted papers). Therefore, our case study can be considered as a *small* graph.

two sets: training and test, both of size $T/2$. For each of the timestamps in the training set, their clustering structures are computed.[3] Then, for each node, they compute the Sociability Index (SoI). Intuitively, the SoI measures the number of times a node changes its cluster along time. For the sake of this paper, the definition of the SoI can be reformulated as follows:

Definition 1 (Sociability Index) The Sociability Index $SoI(x)$ of a node x is:

$$SoI(x) = \begin{cases} \frac{\sum_{t=2}^{T} change(x,t)}{activity(x)} & \text{if } activity(x) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $activity(x) = \sum_{t=2}^{T} active(x,t)$, and the functions $active(x,t)$ and $change(x,t)$ are defined for a node x and a timestamp t as follows:

$$active(x,t) = \begin{cases} 1 & \text{if } x \in V^t \wedge x \in V^{t-1} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$change(x,t) = \begin{cases} 1 & \text{if } active(x,t) \wedge c_t(x) \neq c_{t-1}(x) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

To predict new links, the authors rank the nodes according to their SoI, and remove all nodes below a certain threshold β (in their experiments, $\beta = 0.75$). Additionally, they remove nodes below a node degree threshold ϕ ($\phi = 50$). Then, any pair of two of the remaining nodes that were not clustered together in the training set is a *predicted link*.

In order to measure the performance of the SoI predictor, they check whether these links exist in the test set. They compare this predictor with a random predictor and other methods frequently used in the literature (such as Common Neighbour-based, Adamic-Adar and Jaccard coefficient). The final result is that the link predictor based on the SoI is the best one, and its performance is more than two orders of magnitude greater than the performance of the random predictor.

3.2. The Sociability Index applied to the CCIA network

As stated in the introduction, one of the objectives of this paper is to test whether the SoI is a good predictor also in a small network. To evaluate how this predictor behaves in a small network, we first need to make some considerations.

First, ϕ is very dependent on the structure of the evolutionary graph. In our case, removing nodes with a low node degree is not desirable, since it would result in a graph that is too small. Therefore, instead of filtering nodes by node degree, we use a threshold γ in the *activity* of a node. The activity of a node measures the number of consecutive snapshots containing such node (see Def. 1). Second,

[3] In particular, they use the MCL method for this purpose.

Table 1. Precision of the *SoI*-based and *random* predictors in the CCIA network.

p	20%	20%	20%	30%	30%	30%
γ	1	2	3	1	2	3
precision *random*	0.0014	0.0008	0.0003	0.0012	0.0009	0.0004
precision *SoI*	0.0077	0.0133	0.044	0.0114	0.0253	0.0312

instead of using a threshold β for selecting nodes with highest sociability index, we use a percentage p. Notice that this does not alter the results, but it is useful to compare distinct *SoI*-based predictors, as we shall see in the next section.

To evaluate the performance of a predictor we use the standard *precision* and *recall* metrics, defined as follows:

$$precision = \frac{|Prediction \cap Test|}{|Prediction|}$$

$$recall = \frac{|Prediction \cap Test|}{|Test|}$$

where $Prediction$ is the set of links predicted, and $Test$ is the set of links that actually exist in the test set. While *precision* measures how many of the predicted links are correct, *recall* calculates how many of the existing links were predicted. In many cases, it is useful to have a measure that combines these two values; the F_k-measure was designed to serve this purpose. It can be defined as:

$$F_k = (1 + k^2)\frac{precision \cdot recall}{(k^2 \cdot precision) + recall}$$

Note that the F_k measure always returns a normalized value in $[0, 1]$. Similarly to [2], in our experiments the training set consists of the first $T/2$ timestamps and the test set by the rest. Table 3.2 shows the precision of the SoI when used to predict links in the CCIA network, for different values of p and γ. It can be seen that, even if the performance of the random predictor is improved, the precision is very low. This means that a predicted link has little probability of being correct, something undesirable in link prediction applications.

The CCIA network is a small but rather general conference on AI, in which researchers publish results from different AI areas. Thus, it is unlikely that people who work on different topics—and hence belonging to different clusters—will work together in a publication. Given this assumed behavior, we restrict our prediction to pairs of nodes that have been already clustered together at least once. This allows us to detect pairs of authors working on similar topics. Notice that this information is not explicit in the network. Nevertheless, this restriction is not contradictory to the approach in [2]; it is just an adaptation of the experimental setup for the particular case of a small network. Notice that two authors can have been clustered together without being coauthors of a paper.

Table 3.2 shows the precision, recall and $F_{0.5}$ score for the SoI predictor using $p = \{20\%, 30\%\}$ and $\gamma = \{1, 2, 3\}$ and the restriction discussed above. As it can be seen, increasing the activity threshold γ improves the precision and recall of

Table 2. Effect of activity threshold in SoI precision, recall and $F_{0.5}$ score considering the top 20-30% of nodes with higher SoI.

p	20%	20%	20%	30%	30%	30%
γ	1	2	3	1	2	3
precision	0.065	0.154	0.286	0.112	0.270	0.381
recall	0.152	0.250	0.444	0.333	0.625	0.889
$F_{0.5}$	0.073	0.167	0.308	0.129	0.305	0.430

the SoI predictor; the higher the activity the better the prediction. Furthermore, we can notice that the highest values of precision and recall are achieved with $\gamma = 3$. For instance, with $p = 30\%$ the predictor has precision and recall of 0.38 and 0.89, respectively. Hereinafter, we use an activity threshold $\gamma = 3$.

These results show that the SoI predictor behaves reasonably well in this kind of networks when we predict pairs of nodes that have been clustered together in the past.

4. The Weighted Sociability Index

In this section, we introduce a generalization of the sociability index: the Weighted Sociability Index (SoI_α). As we will show, this index allows to take into account the time in which the organization of nodes into clusters changed.

Definition 2 (Weighted Sociability Index) The Weighted Sociability Index $SoI_\alpha(x)$ of a node x is:

$$SoI_\alpha(x) = \begin{cases} \frac{\sum_{t=2}^{T} change(x,t) \cdot \alpha(t)}{activity(x)} & \text{if } activity(x) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where the functions *activity*, *change* and *active* are defined as before (see Section 3), and α is the time weight function (dependent on t).

Notice that $SoI_\alpha(x)$ is a generalization of the sociability index, since when $\alpha(t) = 1$, i.e., SoI_1, we obtain the original sociability index.

In the following proposition we express the necessary condition that we need to impose on the Weighted Sociability Index in order to take into account the time in which the organization of nodes into clusters changed.

Proposition 1 *Let x and y be two nodes in an evolutionary graph with the same number of cluster changes, and same activity. Also, let x and y be such that their cluster changes can be arranged in pairs in such a way that for every pair, the change for x occured more recently than the one for y. If $\alpha(t)$ is a monotonically increasing function, then $SoI_\alpha(x) > SoI_\alpha(y)$.*

The proof of the previous proposition follows from Definition 2 in a straightforward way. Under the conditions of such proposition, it is easy to see that $SoI_1(x) = SoI_1(y)$, $SoI_t(x) > Sol_t(y)$, and $SoI_{ln(t)}(x) > SoI_{ln(t)}(y)$.

Table 3. Precision, recall and $F_{0.5}$ score of the 4 predictors for different percentage of selected nodes with higher SoI.

	SoI_1			SoI_t			SoI_{ln}		
p	5%	10%	20%	5%	10%	20%	5%	10%	20%
Precision	0.50	0.30	0.28	0.42	0.375	0.32	**0.75**	0.60	0.32
Recall	0.32	0.32	**0.44**	0.32	0.32	**0.44**	0.32	0.32	**0.44**
$F_{0.5}$ score	0.45	0.31	0.31	0.40	0.36	0.35	**0.60**	0.52	0.35

5. Experimental Evaluation

Now, we want to evaluate the performance of predictors based on weighted sociability indexes. First, let us summarize the experimental setup of this evaluation. We test three new predictors, namely the ones using $\alpha(t) = t$, $\alpha(t) = t^2$ and $\alpha(t) = \ln(t)$. We compare them to the predictor based on the original sociability index, i.e., $\alpha(t) = 1$. As before, we rank nodes by their SoI_α indexes, and we also use an activity threshold γ. Additionally, we again focus our prediction on pairs of nodes that were clustered together in the training set. Similarly as in Section 3, we use half of the snapshots for the training set and the rest for the test set.

We perform an exhaustive experimental evaluation. For each predictor, we compute its performance with $p = \{5\%, 10\%, 20\%, 30\%, 50\%\}$ and $\gamma = \{0, 1, 2, 3\}$. We extract the following observations.

First, the performance of the SoI_{t^2} is very similar to the performance of SoI_t (therefore, we do not report the results). We think this is a consequence of the reduced number of timestamps. Second, the performance is better for higher values of γ. Finally, the value of p also has an important impact on the performance. In particular, for values greater than $p = 20\%$, the precision becomes very low, because too many incorrect links are predicted.

In Table 5, we show the most interesting results of the previous experiment. In particular, we only show results when $\gamma = 3$ and $p = \{5\%, 10\%, 20\%\}$. We observe that the precision of the SoI_t and SoI_{ln} predictors are, in general, better than the original SoI_1 predictor, having the best results with SoI_{ln}. The recall, however, remains the same for all predictors. This means that these new predictors are guessing correctly the same number of links, but the number of links predicted is smaller. This is very useful in link predictions where both correct and incorrect predictions matter. Interestingly, the best performance is obtained by SoI_{ln} when $p = 5\%$ and $\gamma = 3$ having a precision of 0.75. This value is exceptionally high in link predictions. Notice that the precision obtained in [2] using SoI_1 is 0.385.

Let us conjecture why this is the case. Both SoI_t and SoI_{ln} give weight to the time in which temporal interactions in the evolutionary graph occurred. However, the difference of weights between recent and old interactions is greater in the first predictor. Therefore, this suggests that, even when both predictors over-perform the one based on the original sociability index, the first interactions still play an important role in the prediction, and hence SoI_{ln} better captures this behavior.

Finally, we perform a prediction based on all available timestamps of CCIA conferences, i.e., using all timestamps as training set, and use the best predictor found in this paper, SoI_{ln}, trained with a $\gamma = 4$ and $p = 5\%$. In Table 5, we

Table 4. Top 10 CCIA authors, ranked by SoI_{ln} (with $p = 5\%$ and $\gamma = 4$).

Author	SoI_{ln}
Antonio Moreno	0.80
Eva Armengol	0.55
Ismael Sanz	0.49
Ramon López de Mántaras	0.40
Aïda Valls	0.37
Cecilio Angulo	0.32
Lledó Museros Cabedo	0.32
Zoe Falomir	0.28
Pilar Dellunde	0.27
Ramón Béjar	0.27

Table 5. Predicted coauthor pairs for the CCIA 2016 conference, obtained with SoI_{ln} with $p = 5\%$ and $\gamma = 4$.

Predicted Pairs		
Antonio Moreno	,	Aïda Valls
Eva Armengol	,	Pilar Dellunde
Ismael Sanz	,	Cecilia Angulo
Ismael Sanz	,	Lledó Museros Cabedo
Ismael Sanz	,	Zoe Falomir
Ramon López de Mántaras	,	Arnau Ramisa
Ramon López de Mántaras	,	Carles Sierra
Cecilio Angulo	,	Lledó Museros Cabedo
Cecilio Angulo	,	Zoe Falomir
Cecilio Angulo	,	Mónica Sánchez
Lledó Museros Cabedo	,	Zoe Falomir
Carles Sierra	,	Maite López-Sánchez

show the top 10 nodes (authors) with highest weighted sociability index. Also, in Table 5, we represent the pairs predicted for the next CCIA edition (CCIA 2016).

6. Related Work

The problem of predicting links with information from interaction networks has been extensively studied in the past two decades. The work in [8] presents a seminal definition of the problem and of the main techniques that can be used to solve it, while a complete survey of existent algorithms and their applications can be found in [9]. These techniques are traditionally designed to analyze networks statically, extracting information from the links that exist in a community in one given moment, ignoring the evolution of networks.

The methods proposed in Asur et al. [2], which we take as basis in this paper, try to overcome this drawback by taking the information extracted from the evolution of the networks for predicting links into account. This information has been proven to be very valuable to analyze the different aspects of a network;

in [1], the authors survey approaches that take it into account to solve problems such as community detection and classification of nodes, in addition to the one of link prediction.

While much of this work aims to develop measures that relate events in time, the particular question of how to weight timed events is, in general, not central. In [12], the authors develop time-aware methods in which events (which are papers in their case study) are considered with different weights according to the time of occurrence. In their approach, the weight of a link between two authors is the time elapsed since the last collaboration between them. A similar approach is presented in [11], where the authors propose to increase the importance of recent events by including a logarithmic function over time in the computation of a proximity score between two nodes.

Our work proposes measures to extract dynamic information of networks that evolve in time, using a weighting technique to give more importance to recent events. Differently to the approaches discussed in the last paragraph, we take into account evolution in time to analize properties of one particular node in a network (such as the Sociability Index), instead of purely predicting relations between nodes.

Finally, Garcia-Gasulla et al. [7] present a study of the link prediction problem for large graphs. In that case, they are mainly concerned with the efficient processing of huge amounts of data. Dealing with small networks presents different challenges. Since large quantities of data, from which information can be extracted, are not available, the particular structure of the community needs to be exploited.

7. Conclusions and Future Work

The contribution of this paper is two-folded. On the one hand, we adapt the methods for predicting links proposed in [2] to the case study of the CCIA coauthorship network. This evolutionary graph has some particularities that make some predictors extensively used in the literature inaccurate. This is the case of the predictor based on the sociability index. The bad performance of this predictor can be explained as follows. First, it is a small community, and therefore the dataset is reduced, both for training and for testing. Second, it is a general conference, with participants who work in many different sub-topics from Artificial Intelligence. This makes new collaborations between members less likely than in communities where members work in the same area. Finally, being a regional community, some people can be very active during a period of time and then stop participating, e.g., changing their affiliation. To overcome these drawbacks in the prediction, we modified the predictors adapting them to the topology of the network we study. Specifically, we predict only links between people who have already shared a cluster, representing the natural division by topics.

On the other hand, to take into account the activity pattern, we developed weighted sociability indexes. They give more relevance to the events that happened more recently in time. Both changes result into noticeable improvements in the precision of the predictions. In particular, we achieved a precision of 0.75 using one of our new predictors.

As future work, we plan an exhaustive evaluation of our time-aware predictors in large evolutionary networks. Another interesting direction of research is the use of these approaches in order to predict new papers. Notice that now we are only predicting pairs of co-authors rather than the set of authors of a paper. A possible extension to do so is the use of hypergraphs, where each paper is represented by a hypernode. This way, the problem is reduced to the detection of new hypernodes (and their corresponding hyperedges).

Acknowledgements

This work is partially funded by the MINECO/FEDER project RASO (TIN2015-71799-C2-1-P), the CSIC projects 201450E045 and 201550E022, the COINVENT project (FET-Open grant number: 611553) and ESSENCE project (ITN 607062). Jesús Giráldez-Cru acknowledges the IIIA-CSIC, where this research was carried out. The authors thank the anonymous reviewers for their valuable comments and Marc Pujol for his insightful comments.

References

[1] C. Aggarwal and K. Subbian. Evolutionary network analysis: A survey. *ACM Comput. Surv.*, **47**(1)(2014), 1–36.

[2] S. Asur, S. Parthasarathy, and D. Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. *ACM Trans. Knowl. Discov. Data*, **3**(4)(2009), 1–36.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008** (10)(2008).

[4] J. G. Cleary and L. E. Trigg. K*: An instance-based learner using an entropic distance measure. In *Proc. of ML'95*, pages 108–114, 1995.

[5] S. Dongen. A cluster algorithm for graphs. Technical report, 2000.

[6] S. Fortunato. Community detection in graphs. *Physics Reports*, **486**(3-5)(2010), 75–174.

[7] D. Garcia-Gasulla, C. Cortes, E. Ayguade, and J. Labarta. Evaluating link prediction on large graphs. In *Artificial Intelligence Research and Development: Proceedings of the 18th International Conference of the Catalan Association for Artificial Intelligence*, pages 90–99. IOS Press, 2015.

[8] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, **58**(7)(2007), 1019–1031.

[9] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, **390**(6)(2011), 1150–1170.

[10] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, **69**(2)(2004).

[11] P. R. Soares and R. B. Prudêncio. Proximity measures for link prediction based on temporal events. *Expert Systems with Applications*, **40**(16)(2013), 6652–6660.

[12] T. Tylenda, R. Angelova, and S. Bedathur. Towards time-aware link prediction in evolving social networks. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, SNA-KDD '09, pages 9:1–9:10, New York, NY, USA, 2009. ACM.