Journal of Computational Science 13 (2016) 1–9

Contents lists available at ScienceDirect

Journal of Computational Science

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

Adaptive degree penalization for link prediction

Víctor Martínez*, Fernando Berzal, Juan-Carlos Cubero

CITIC & Department of Computer Science and Artificial Intelligence, University of Granada, Spain

ARTICLE INFO

Article history:
Received 23 December 2014
Received in revised form 12 June 2015
Accepted 10 December 2015
Available online 24 December 2015

Keywords:
Link prediction
Networks
Graphs
Topology
Shared neighbors

ABSTRACT

Many systems of interest are best described using networks that represent binary relationships among their elements. Link prediction aims to infer the link formation process by predicting missed or future relationships based on currently observed connections. Different techniques and measures have been proposed in the literature to solve this problem. Similarity-based local methods achieve high precision with a low computational complexity. However, determining which particular technique should be applied for each particular network remains an open question. In this paper, we exploit the existence of a relationship between the best-performing degree of penalization for shared neighbors and the network clustering coefficient. We propose an adaptive degree penalization link prediction method, a novel link prediction technique that achieves better results than previously proposed methods.

1. Introduction

The link prediction problem consists of inferring the formation of new relationships or the existence of still-unknown connections between pairs of entities in a network based on their properties and currently observed links [24]. This problem has attracted a lot of attention, since a large number of systems in many different fields can be described using networks. Approaches and techniques to solve this problem allow us to extract implicit information present in the network, identify spurious links, or model and evaluate network evolution mechanisms. These problems are of great interest since they are closely related to other problems usually found in different disciplines. For example, link prediction has been used to predict previously unknown protein interactions in protein-protein interaction networks [27]. It has also been used to study and predict future author collaborations and tendencies in co-authorship networks [34]. In fact, link prediction is present in our daily lives when we get friendship suggestions in social networks [10] or recommendations of new products in e-commerce web sites [16].

The link prediction problem is formally defined as follows. Let G be an undirected graph $G=(V, E)$, where V is a set of optionally labeled nodes and E is a set of edges (also referred to as links) between pairs of elements from set V. Given a snapshot of the network G at time t, the link prediction problem consists of inferring the subset of missing links in the current snapshot that will be formed at time $t+\Delta$.

Some notational conventions are important to properly describe the proposed solutions for the link prediction problem. An edge between nodes x and y is denoted as $e_{x,y}$. The number of nodes in the network is $|V|$. The number of edges is $|E|$. The set of nodes connected through an edge to a node x is called the neighborhood of x and is referred to as Γ_x. The degree of a node x in an undirected graph is defined as the number of edges connected to the node and will be denoted as $|\Gamma_x|$.

Many link prediction methods are based on the observation that nodes that share a higher number of neighbors are more likely to be connected [30]. Well-known link prediction techniques take into account the number of directly shared neighbors (local methods) or the number of chains of neighbors between two nodes (global methods) to estimate the probability of the existence of a potential link. However, these techniques always work the same way, regardless of the network they are applied to. Our work is motivated by the lack of further studies about how link prediction techniques are affected by network structural properties and how existing methods can be adapted to the structural properties of particular networks in order to obtain better results.

This paper is organized as follows. Related work is presented in Section 2. We propose and describe a generalized degree penalization similarity measure in Section 3. In Section 4, we analyze the relationship of the best-performing degree penalization with respect to the topological properties of the network. A novel link prediction technique called adaptive degree penalization is presented in Section 5. Finally, the conclusions drawn from this study and some lines of future research are presented in Section 6.

* Corresponding author.
E-mail address: victormg@acm.org (V. Martínez).

http://dx.doi.org/10.1016/j.jocs.2015.12.003
1877-7503/

2. Related work

Link prediction has been the subject of many studies [12]. A large number of techniques following different approaches have been proposed to deal with the link prediction problem [26]. In this work, we limit our scope to techniques that consider only network topology, albeit methods considering other attributes have also been proposed [13].

The first and most studied approach is based on the similarity between nodes [24,26]. Similarity-based techniques assume that nodes are more likely to form links with similar nodes. A function that assigns a similarity score $s(x, y)$ to every pair of nodes in the network is defined. This similarity score can take into account different features, which can be topological properties or network-specific attributes. All possible pairs of nodes are ranked in decreasing order based on their similarity scores. Links at the top of the ranked list are supposed to be more likely to be present in the set of missing links.

Similarity-based methods can be categorized depending on the amount of information taken into consideration when computing the similarity function. For example, local similarity techniques consider only direct neighbor information. This family of techniques can achieve high precision in most networks and have a linear time complexity, which makes them suitable for large networks. On the other hand, global methods use the whole topology of the network to compute the similarity score for every possible link. This type of techniques has the advantage of being able to compute the similarity between each pair of nodes regardless of their distance within the network, instead of being limited to neighbor-sharing pairs of nodes. Their main drawbacks are their high computational complexity and their sensitivity to noise, which usually leads to lower precision than local methods. Finally, quasi-local techniques have been proposed to try to find an equilibrium between the amount of considered information and the computational complexity of the resulting methods. Most quasi-local techniques are either based on local ones with small variations to consider neighbors of neighbors or based on global ones with constraints on the lengths of the considered paths.

An alternative approach is to describe the network formation model in statistical terms. Statistical approaches build a parameterized model assuming the existence of a known structure in the network [15,42,7,14]. The parameters of the model for a particular network are estimated using statistical methods. Finally, the adjusted model is used to compute the probability of the formation of each possible link. The main problem of this kind of techniques is that they suffer from a very high computational cost, which limits their applicability to networks of only hundreds or a few thousand nodes. In addition, they can only be applied to networks with a particular structure.

Other algorithmic approaches have also been proposed. Since link prediction techniques are inherently heuristic, some metaheuristic-based methods have been proposed in order to automatically adjust the influence of a set of local similarity-based techniques in an attempt to maximize precision [5]. The link prediction problem can be seen as a classification problem with two classes (existence and absence of links). This point of view allows the application of traditional machine learning techniques [19,8]. These techniques can obtain better results than other approaches at the cost of a previous training stage, which is not always possible in many applications. Furthermore, they have the drawback that the predictive model they build is often hard to understand and analyze.

In this paper, we focus our attention on local similarity-based techniques, since these techniques are widely used due to their high scalability and the reasonable precision they obtain [45]. Computational complexity is really important in link prediction, since most real networks are huge, with hundreds of thousands or even millions of nodes. Even worse, there are usually time or resource constraints in many problems related to link prediction. In fact, most recommender systems use only local techniques.

The most basic local method is called Common Neighbors (CN). This technique assigns a score based just on the number of shared neighbors:

$$s^{CN}(x, y) = |\Gamma_x \cap \Gamma_y| \tag{1}$$

It makes sense to assume that, if two individuals share many acquaintances, they are more likely to meet than two individuals without common contacts. Different studies have confirmed this hypothesis by observing a correlation between the number of shared neighbors between pairs of nodes and their probability of being linked [30].

Lada Adamic and Eytan Adar proposed the Adamic-Adar Index (AA) to measure the similarity between two entities based on their shared features [1]. This measure was adapted to link prediction by considering shared neighbors as features:

$$s^{AA}(x, y) = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{\log |\Gamma_z|} \tag{2}$$

This equation is a variation of the common neighbors similarity function. Here, each shared neighbor is penalized by its degree. This intuitively makes sense in a large number of real-world networks. For example, in social networks, the amount of resources or time that a node can spend on each of its neighbors decreases as its degree increases, also decreasing its influence on them.

The Resource Allocation Index (RA) was motivated by the resource allocation process which takes place in complex distribution networks [45]. It models the transmission of resources between two unconnected nodes x and y through neighborhood nodes. Each neighborhood node gets a given amount of resources and distributes them evenly among its neighbors. The amount of resources obtained from node x by node y through their shared neighbors can be considered as a similarity measure between both nodes. The resource allocation index has shown to be the local measure with better results in a large number of networks [25]. It can be computed as

$$s^{RA}(x, y) = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{|\Gamma_z|} \tag{3}$$

Other local similarity-based techniques have been proposed, including the Preferential Attachment Index [3], the Jaccard Index [18], the Salton Index [37], the Sørensen Index [40], the Hub-Promoted and Hub-Depressed Indices [36], and the Leicht-Holme-Newman Index [22]. Most of these techniques are variations of the previously described measures, but also consider other features such as the number of unshared neighbors. Different comparative studies have shown that these variations work better in very specific contexts, yet are worse on average [45].

3. Similarity based on adjustable degree penalization

The Common Neighbors method, the Adamic-Adar Index, and the Resource Allocation Index have been presented in the literature as three different link prediction techniques. It can be readily seen that these methods assume that the probability of existence of a link between two nodes is proportional to the number of shared neighbors between them, but penalize each one according to their degree, with a null penalization in the Common Neighbors case. From our point of view, CN, AA, and RA are just variations of the same technique, which considers shared neighbors and penalizes them by their degree using different penalization schemes.

In addition, these techniques limit the degree of penalization to fixed values without taking into account that the most suitable penalization degree could be different according to the peculiarities of the network under study.

A generalized expression for the existing degree penalization local link prediction techniques that allows us to specify the desired degree penalization level can be stated as

$$s^{GDP}(x,y) = \sum_{z \in \Gamma_x \cap \Gamma_y} |\Gamma_z|^{-\alpha} \quad (4)$$

where α is a free parameter to adjust the penalization to each particular network and $|\Gamma_z|$ is the degree of the shared neighbor z. It can be seen that this Generalized Degree Penalization (GDP) expression is equivalent to the Common Neighbors method when $\alpha = 0$, when no penalization is performed. Furthermore, this expression is equal to the Resource Allocation Index when $\alpha = 1$. Finally, the behavior of the Adamic-Adar Index can be closely approximated by setting α to a value between 0 and 1. For example, we obtained $\alpha \approx 0.37$ after fitting the function $1/x^{\alpha}$ to the function $1/\log x$ in the interval [2, 20], which encompasses the expected degree values for most nodes in typical real-world networks.

Our definition of local similarity offers two main benefits. First, it allows us to unify the analysis of three existing measures instead of having three different related methods. Second, existing measures do not obtain optimal results since the optimal value of the α parameter is not adjusted by existing techniques.

Which one of those three variations works better in practice depends upon the network they are applied to. However, no advances in determining which technique is better have been accomplished. A complete empirical evaluation is typically done on a case-by-case basis. The ideal degree penalization scheme varies among networks, since each of the aforementioned local techniques obtains better results than the others depending on the network under analysis. Since those techniques only rely on topological properties, the optimal α value should be determined by the network structure.

In order to understand how different values of α behave in different networks with different properties, we tested the α parameter for a reasonable range of values and plotted the precision and the AUC obtained for each value, as shown in Figs. 1 and 2, respectively. The best-performing degree penalization was estimated for a collection of networks, which we describe in the appendix, by applying the Generalized Degree Penalization technique to each network and varying α in a range from -1.0 to 2.0 in steps of size 0.1. We measured the obtained precision and AUC for each network by performing a 5-fold cross-validation experiment as described in the appendix.

Our experiments show that the best-performing α value can reasonably vary for different networks. It can be observed that the α values, fixed by CN, AA, and RA, are not always even near to the best-performing α. Hence, a technique to adapt the α parameter to the network would be desirable. As far as we know, there are no previous studies about the best-performing degree penalization relationship to structural network properties. The only step that has been taken in this direction is the observation that a higher performance can be obtained by increasing or decreasing the degree penalization depending on the node degree [41]. However, [41] suffers from the same limitations as CN, AA, and RA, since the particular values used for degree penalization are also fixed for every network.

4. Relating degree penalization to the network structural properties

The best-performing α value for each network might depend on the network structure. Intuitively, you might conjecture that some topological properties of the network might be related to this value. Guided by this conjecture, we carried out an experiment to find the degree of correlation between the best-performing value of α and different quantifiable global topological properties of networks, whose results can be found below.

Table 1
Correlation coefficients (and Bonferroni-adjusted p-values) between network structural properties and the best-performing alpha value according to precision and AUC in our link prediction experiments. Properties, from top to bottom: number of nodes ($|V|$), number of edges ($|E|$), average degree ($\langle k \rangle$), average clustering coefficient (C), average shortest path length ($ASPL$), diameter (D), heterogeneity (H), and assortativity (r).

Property	Coefficients for precision	Coefficients for AUC
$\|V\|$	0.1477 (1.0)	−0.0553 (1.0)
$\|E\|$	0.5980 (0.2966)	0.4288 (1.0)
$\langle k \rangle$	0.6041 (0.2731)	0.6498 (0.1397)
C	0.9374 (4.06×10^{-6})	0.8812 (0.0002)
$ASPL$	−0.4990 (0.9325)	−0.4205 (1.0)
D	−0.4203 (1.0)	−0.3758 (1.0)
H	−0.3425 (1.0)	−0.4250 (1.0)
r	0.4975 (0.9466)	0.4289 (1.0)

We took the best-performing α value for each network and computed the Pearson correlation coefficient against different network properties to find potential correlations between the best-performing α value and the network structural properties. The value of each network topological property was computed for each training network generated in the 5-fold cross validation process used in our experiments, as described in the appendix. The obtained correlation coefficients are shown in Table 1.

As expected, some structural network properties are slightly correlated to the best performing-degree penalization value. For example, the assortativity is weakly correlated to this value. This implies that the best-performing α tends to be higher as the nodes of the network tend to be connected to similar ones in terms of their degree. On the other hand, the average shortest path length shows a negative correlation. This suggests that the best-performing α tends to increase as the length of the shortest paths between nodes decreases. However, these correlations are weak and do not help us predict the α value that should be used to improve precision or AUC in link prediction.

On the other hand, the obtained results show that the average clustering coefficient of a network and the best-performing α are strongly correlated ($r = 0.9374$ for precision and $r = 0.8812$ for AUC). The average clustering coefficients have been plotted against the best-performing alpha values for precision in Fig. 3 and for AUC in Fig. 4. It can be seen that they follow an almost linear correlation. This implies that the best-performing α value can be estimated with a reasonable accuracy by attending just to the average clustering coefficient measured in the network.

As far as we know, this relationship has not been previously mentioned nor documented. The discovered relationship is coherent with previous results and allows us to explain why the Common Neighbors method usually obtains better results in low-clustered networks and the Resource Allocation method tends to obtain better results in highly clustered ones.

5. Adaptive degree penalization

The generalized degree penalization measure, combined with the observation that the degree penalization that obtains better results in link prediction can be estimated by considering the clustering coefficient of the network, allows us to propose a new link prediction technique that automatically adapts to the network.

We propose an adaptive degree penalization link prediction technique whose definition of local similarity tries to estimate the best-performing degree penalization by using the average

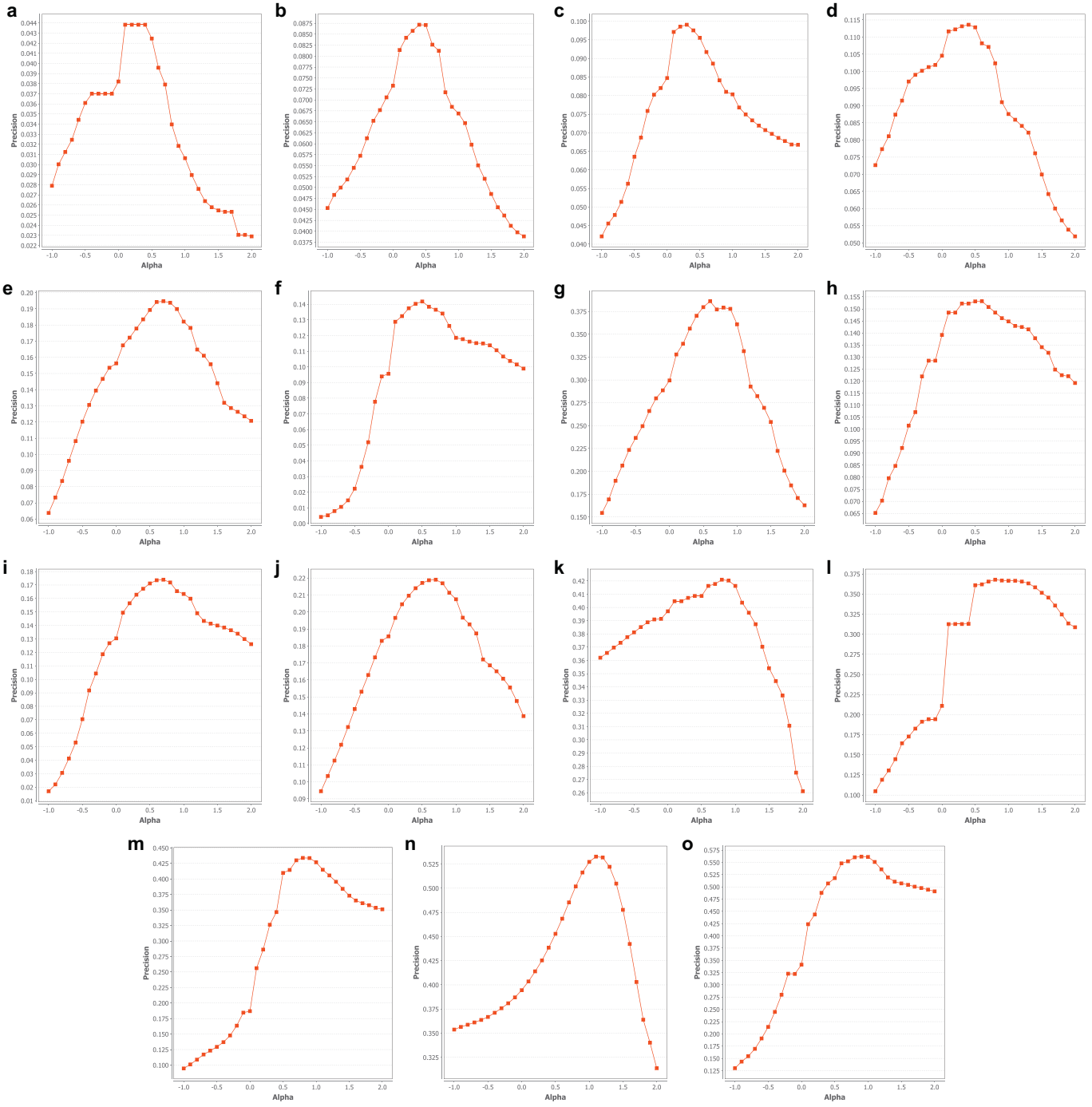

Fig. 1. Precision obtained by our link predictor varying the alpha parameter for different networks (see appendix for a description of the networks used in our experiments).

clustering coefficient observed in the network. We define our similarity measure as

$$s^{ADP}(x,y) = \sum_{z \in \Gamma_x \cap \Gamma_y} |\Gamma_z|^{-\beta C} \tag{5}$$

where C is the average clustering coefficient of the network and β is a constant. The average clustering coefficient only has to be measured once before applying our link prediction algorithm. In really large or very dynamic networks, it could be estimated by a sampling procedure. The β constant is determined beforehand using an heterogeneous set of networks, since all the networks we have tested seem to follow the same correlation pattern between the clustering coefficient and the best-performing degree penalization.

In order to test the performance of the proposed technique, we have carried out an experiment using an estimated β value. To determine this value, we performed a linear regression between the clustering coefficients of a set of networks, which we call training networks, and the best-performing alpha values. We obtained an slope of $\beta = 2.52$ for precision and a slope of $\beta = 2.47$ for AUC. Since the slope is consistent for both measures, we set the estimated slope to be $\beta = 2.5$.

Using this value for the β parameter, we drove a 5-fold cross validation experiment to determine the precision and AUC of the proposed link prediction method. This experimentation is carried

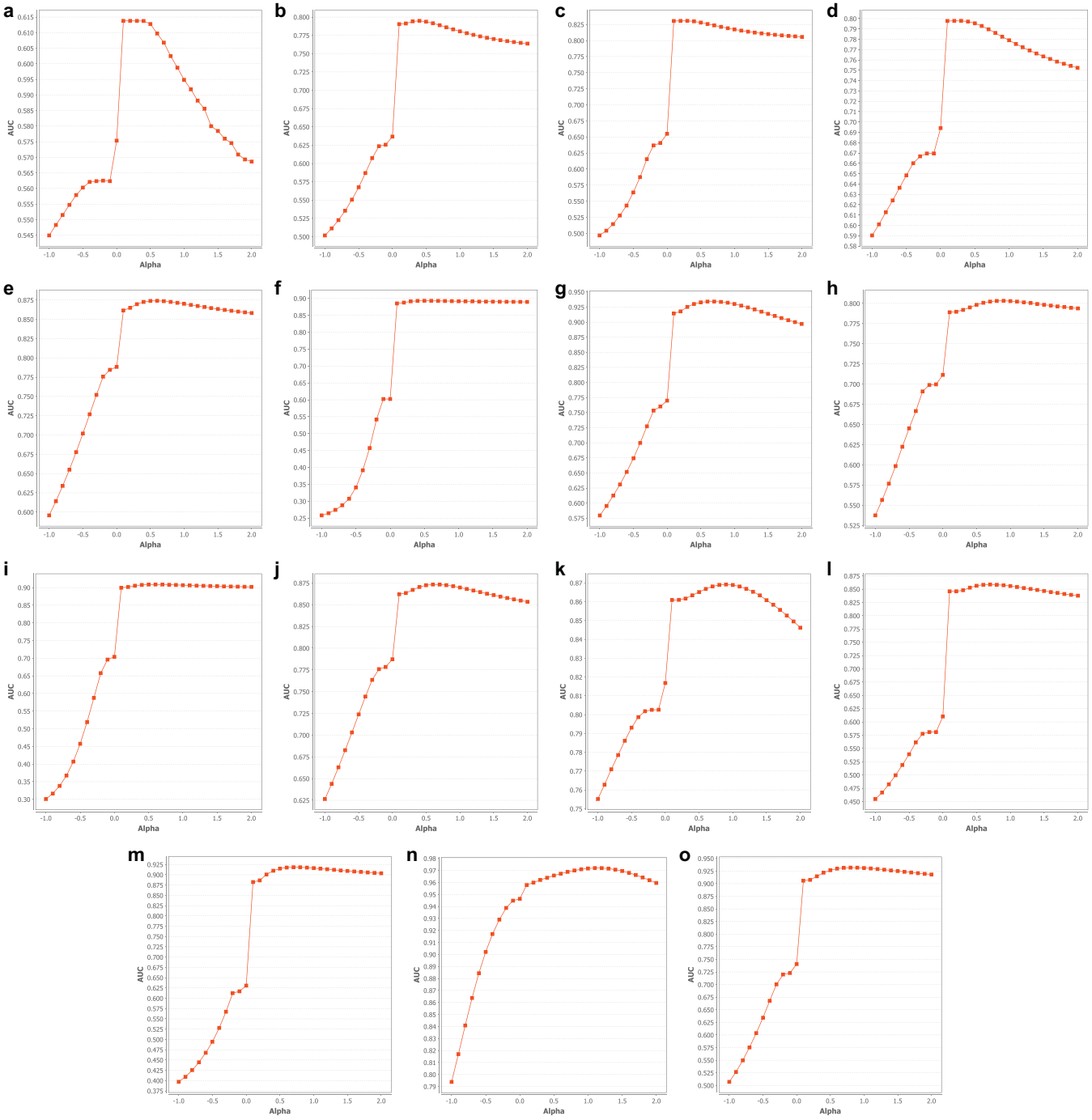

Fig. 2. AUC obtained by our link predictor varying the alpha parameter for different networks (see appendix for a description of the networks used in our experiments).

out for the training set (networks used to determine the optimal β) and for a test set (networks not involved in the β estimation). We compared our method precision and AUC against the results obtained by the Common Neighbors method, the Adamic-Adar Index, and the Resource Allocation method. In addition, we also compared it with the best precision and AUC obtained by varying the α value. It must be taken into account that, since the best precision has been estimated with a step resolution of size 0.1, our β-based technique could achieve slightly better results than the estimated best-performing α. In fact, the precision obtained by adaptive degree penalization is even higher than that determined by varying the α parameter in the FBK network, and the AUC is higher in the SMG and BUP networks.

The results that we obtained in our experiments are shown in Table 2 for precision and Table 3 for AUC. It can be seen how our adaptive degree penalization method obtained the best results for precision in 18 of the 22 network when compared to CN, AA, and RA (13 from training set and 5 from test set). With respect to AUC, our method obtained the best results for 18 of the 22 networks (12 from training set and 6 from test set). In most of the remaining cases, it still stands as the second best method. On average, our method obtains a higher precision of 0.0082 than the Resource Allocation method, the best of the previously proposed local techniques. This improvement may seem small. However, it can be seen that the best precision based on varying α for each network is only 0.0026 better than the average precision obtained by our ADP method.

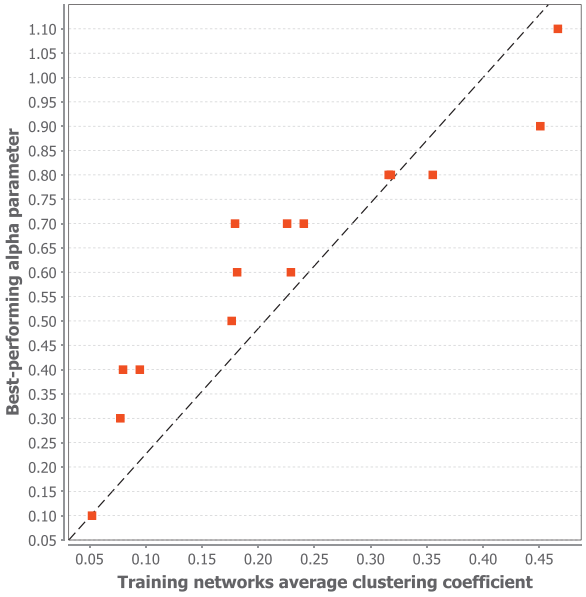

Fig. 3. Average clustering coefficient shown against the best-performing alpha value according to precision for the networks used in our experiments.

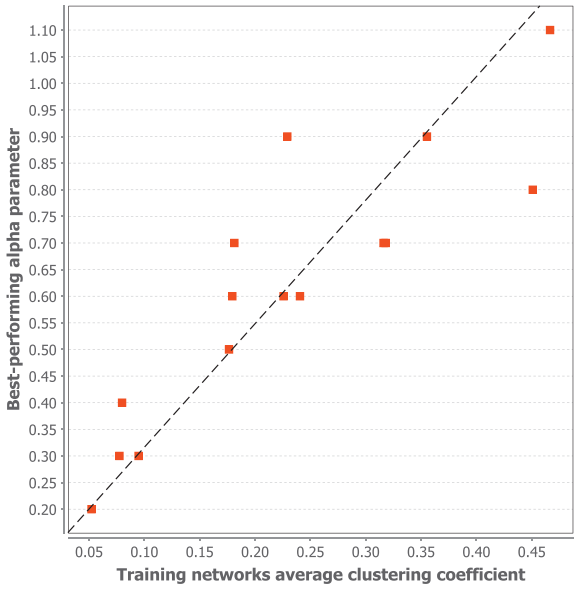

Fig. 4. Average clustering coefficient shown against the best-performing alpha value according to AUC for the networks used in our experiments.

Considering AUC, ADP is also better than the best existing method, RA in this case, and almost indistinguishable from the best result obtained by testing for multiple values of α.

We performed a Friedman test [11] to determine if there are statistically significant differences among the link prediction methods used in our experiments. The Friedman tests confirm that there are significant differences for precision and AUC, since the obtained p-values are 2.5571×10^{-8} and 1.1921×10^{-11}, respectively. The average ranks obtained in our experiments by the different methods for precision were: 1.36 for CN, 2.55 for AA, 2.34 for RA, 3.75 for our method (ADP). For AUC, the ranks were 1.00 for CN, 2.36 for AA, 2.84 for RA, 3.80 for ADP.

Finally, we performed a post-hoc test using the Wilcoxon signed-rank test [44] with our link prediction method as control. The alpha level (initially set to 0.05) was adjusted with the Holm method to control the familywise error rate (FWER). According to precision, when we compared our method to CN, AA, and RA, we obtained 8.8219×10^{-4}, 0.0034, and 0.0198 as corrected p-values, respectively. According to AUC, the corrected p-values were 8.8219×10^{-4}, 0.0020, and 0.0461. In other words, our method obtains significantly better results than previous methods. In addition, we obtained a corrected p-value of 0.0055 for precision and 0.0801 for AUC when we compared our adaptive degree penalization method to the method that tests many different α values. This result indicates that the test failed to reject the null hypothesis for

Table 2
Results obtained from our method comparison. Columns, from left to right: network name, average clustering coefficient of the training network, estimated α by $\beta\hat{C}$, best-performing α value from experiment (see Fig. 1), precision obtained by the best-performing α, Common Neighbors precision, Adamic-Adar Index precision, Resource Allocation precision, and adaptive degree penalization precision.

	Network	$\hat{C}$	Estimated α	Best α	Best precision	CN	AA	RA	ADP
Training networks	UPG	0.05	0.13	0.1	0.0438	0.0411	0.0331	0.0306	**0.0438**
	HPD	0.08	0.20	0.4	0.0872	0.0748	0.0828	0.0669	**0.0842**
	ERD	0.08	0.19	0.3	0.0991	0.0883	0.0950	0.0803	**0.0986**
	YST	0.09	0.24	0.4	0.1136	0.1103	0.1080	0.0876	**0.1127**
	ADV	0.18	0.45	0.7	0.1947	0.1591	0.1783	0.1821	**0.1862**
	KHN	0.18	0.44	0.5	0.1418	0.1085	0.1382	0.1185	**0.1407**
	PGP	0.18	0.45	0.6	0.3861	0.3058	0.3655	0.3608	**0.3761**
	CEG	0.23	0.57	0.6	0.1532	0.1401	**0.1532**	0.1448	0.1527
	LDG	0.23	0.56	0.7	0.1740	0.1361	0.1662	0.1634	**0.1727**
	ZWL	0.24	0.60	0.7	0.2190	0.1891	0.2110	0.2075	**0.2188**
	INF	0.36	0.89	0.8	0.4210	0.3978	0.4080	0.4163	**0.4192**
	HTC	0.32	0.80	0.8	0.3679	0.2406	0.3583	0.3667	**0.3673**
	CGS	0.32	0.79	0.8	0.4339	0.2037	0.3986	0.4265	**0.4337**
	FBK	0.47	1.17	1.1	0.5327	0.3946	0.4185	0.5272	**0.5334**
	CDM	0.45	1.13	0.9	0.5617	0.3878	0.5338	**0.5611**	0.5466
Test networks	EML	0.16	0.41	0.6	0.1990	0.1825	0.1923	0.1802	**0.1967**
	SMG	0.22	0.55	0.4	0.1611	0.1420	0.1587	0.1408	**0.1593**
	BUP	0.37	0.94	0.7	0.2608	0.2449	**0.2608**	0.2563	0.2540
	GRQ	0.36	0.90	0.8	0.5550	0.4002	0.5308	0.5531	**0.5539**
	HMT	0.40	1.00	0.9	0.3977	0.2580	0.3267	**0.3959**	**0.3959**
	UAL	0.46	1.15	1.0	0.5160	0.4388	0.4558	**0.5160**	0.5155
	NSC	0.47	1.18	1.2	0.6667	0.5058	0.6295	0.6659	**0.6667**
	Training set average				0.2620	0.1985	0.2432	0.2494	**0.2591**
	Test set average				0.3938	0.3103	0.3649	0.3869	**0.3917**
	Overall average				0.3039	0.2341	0.2820	0.2931	**0.3013**

Bolded values indicate the best performance among the compared methods (columns) per network (rows).

Table 3
Results obtained from our method comparison. Columns, from left to right: network name, average clustering coefficient of the training network, estimated α by $\beta\hat{C}$, best-performing α value from experiment (see Fig. 2), AUC obtained by the best-performing α, Common Neighbors AUC, Adamic-Adar Index AUC, Resource Allocation AUC, and adaptive degree penalization AUC.

	Network	$\hat{C}$	Estimated α	Best α	Best AUC	CN	AA	RA	ADP
Training networks	UPG	0.05	0.13	0.2	0.6138	0.5882	0.6036	0.5944	**0.6137**
	HPD	0.08	0.20	0.4	0.7948	0.7080	**0.7941**	0.7805	0.7912
	ERD	0.08	0.19	0.3	0.8307	0.7355	0.8305	0.8172	**0.8306**
	YST	0.09	0.24	0.3	0.7978	0.7336	0.7961	0.7790	**0.7978**
	ADV	0.18	0.45	0.6	0.8741	0.8231	0.8683	0.8701	**0.8734**
	KHN	0.18	0.44	0.5	0.8929	0.7436	0.8909	0.8916	**0.8928**
	PGP	0.18	0.45	0.7	0.9342	0.8373	0.9288	0.9303	**0.9318**
	CEG	0.23	0.57	0.9	0.8033	0.7443	0.7931	**0.8029**	0.8002
	LDG	0.23	0.56	0.6	0.9088	0.7978	0.9053	0.9069	**0.9088**
	ZWL	0.24	0.60	0.6	0.8733	0.8202	0.8677	0.8698	**0.8733**
	INF	0.36	0.89	0.9	0.8692	0.8315	0.8642	0.8689	**0.8691**
	HTC	0.32	0.80	0.7	0.8587	0.7134	0.8579	0.8558	**0.8584**
	CGS	0.32	0.79	0.7	0.9184	0.7495	0.9123	0.9162	**0.9182**
	FBK	0.47	1.17	1.1	0.9720	0.9514	0.9615	0.9716	**0.9720**
	CDM	0.45	1.13	0.8	0.9318	0.8143	0.9243	**0.9311**	0.9298
Test networks	EML	0.16	0.41	0.5	0.7976	0.7561	0.7970	0.7885	**0.7974**
	SMG	0.22	0.55	0.5	0.8331	0.7535	0.8296	0.8290	**0.8332**
	BUP	0.37	0.94	1.1	0.7486	0.6977	0.7395	0.7486	**0.7489**
	GRQ	0.36	0.90	0.9	0.9259	0.8073	0.9171	0.9257	**0.9259**
	HMT	0.40	1.00	0.9	0.9115	0.8417	0.8954	**0.9112**	**0.9112**
	UAL	0.46	1.15	1.1	0.9313	0.8741	0.9141	0.9310	**0.9313**
	NSC	0.47	1.18	1.0	0.9374	0.8037	0.9327	**0.9374**	0.9366
	Training set average				0.8583	0.7728	0.8532	0.8524	**0.8574**
	Test set average				0.8693	0.7906	0.8608	0.8673	**0.8692**
	Overall average				0.8618	0.7785	0.8556	0.8572	**0.8612**

Bolded values indicate the best performance among the compared methods (columns) per network (rows).

AUC, so there is no evidence to suggest that the AUC obtained by our adaptive degree penalization method is significantly different from the precision obtained by an exhaustive search of possible values for the α penalization parameter.

6. Conclusions and future work

In this paper, we have presented a generalized adaptive degree penalization link prediction method that unifies previously proposed local similarity-based techniques. We have also studied how the penalization parameter relates to some network structural properties, finding a strong correlation between the clustering coefficient and the best-performing value of the degree penalization parameter. This result allowed us to propose a new technique that automatically estimates this parameter based on the clustering coefficient measured in the network under study. Our new method obtains statistically significant better results than CN, AA, and RA in an experimental evaluation considering 22 networks from different application domains.

These results lead to a better understanding of local similarity-based techniques. Instead of considering different measures with different theoretical backgrounds, we consider them as instances of a more general technique. In addition, our observation that a higher degree penalization should be performed as the average clustering coefficient of the network increases is consistent with known results.

A further theoretical study of these empirical results remains as future work. A pure theoretical study could lead to interesting results. Some studies have started to try to understand the behavior of the clustering coefficient in graphs generated from network models [9,39]. Connecting network formation and link prediction theory could help improve the results obtained by link prediction techniques.

Acknowledgements

Work partially supported by the Spanish Ministry of Economy and the European Regional Development Fund (FEDER), under grant TIN2012-36951, and partially by the Ministry of Education of Spain under the program “Ayudas para contratos predoctorales para la formación de doctores 2013” (grant BES-2013-064699).

Appendix A.

A.1. Network structural measures

Different structural network measures were considered to summarize network topologies. These measures help us understand networks at a macroscopic level. Different measures related to shortest paths were taken into account. For example, we computed the average shortest path length (ASPL) considering every possible pair of nodes. In addition, we also obtained the network diameter (D), which is the length of the longest shortest path in the network.

We also computed the clustering coefficient [43], which measures the tendency of a node to cluster with other nodes forming triangles. More intuitively, in a social network, the clustering coefficient would measure the tendency of the friends of a given person to be also friends among themselves. We computed the average clustering coefficient of the network as

$$\bar{C} = \frac{\sum_{x \in V} C_x}{|V|} \tag{6}$$

where C_x is the clustering coefficient, ranging from 0 to 1, of a node x, which is itself computed as

$$C_x = \frac{|\{e_{y,z} : y \in \Gamma_x, z \in \Gamma_x, e_{y,z} \in E\}|}{|\Gamma_x|(|\Gamma_x| - 1)}. \tag{7}$$

Heterogeneity is another interesting measure that is related to the node degree distribution of the network [45]. It measures the variance of node degrees. In a typical real-world network, a higher value for heterogeneity represents a larger number of high-degree nodes compared to the number of low-degree nodes. This value can be computed as

Table 4
Network topological properties (with networks sorted by their clustering coefficient). Columns, from left to right: network name, number of nodes ($|V|$), number of edges ($|E|$), average degree ($\langle k \rangle$), average clustering coefficient (C), average shortest path length ($ASPL$), diameter (D), heterogeneity (H), and assortativity (r).

Name	$\|V\|$	$\|E\|$	$\langle k \rangle$	C	$ASPL$	D	H	r	Reference
UPG	4941	6594	2.67	0.08	18.99	46	1.4504	0.0035	[43]
HPD	8756	32331	7.38	0.11	4.19	14	4.5133	−0.051	[35]
ERD	6927	11850	3.42	0.12	3.78	4	12.6708	−0.1156	[4]
YST	2284	6646	5.82	0.13	4.29	11	2.8479	−0.0991	[6]
EML	1133	5451	9.62	0.22	3.61	8	1.9421	0.0782	[6]
ADV	5155	39285	15.24	0.25	3.22	9	5.4060	−0.0951	[28]
KHN	3772	12718	6.74	0.25	3.63	12	9.422	−0.1205	[4]
PGP	10680	24316	4.55	0.27	7.49	24	4.1465	0.2382	[38]
CEG	297	2148	14.46	0.29	2.46	5	1.8008	−0.1632	[43]
LDG	8324	41532	9.98	0.31	4.37	16	6.188	−0.0997	[4]
SMG	1024	4916	9.6	0.31	2.98	6	3.9475	−0.1925	[4]
ZWL	6651	54182	16.29	0.32	3.85	10	2.5851	0.0006	[4]
INF	410	2765	13.49	0.46	3.63	9	1.3876	0.2258	[17]
BUP	105	441	8.4	0.49	3.08	7	1.4207	−0.1279	[20]
HTC	7610	15751	4.14	0.49	5.68	19	2.0986	0.2939	[31]
CGS	6158	11898	3.86	0.49	3.62	14	3.9467	0.2426	[4]
GRQ	5241	14484	5.53	0.53	5.05	17	3.0523	0.6593	[23]
HMT	2426	16630	13.71	0.54	3.15	10	3.1011	0.0474	[21]
FBK	4024	87887	43.68	0.59	3.98	13	2.432	0.0707	[29]
UAL	332	2126	12.81	0.63	2.74	6	3.4639	−0.2079	[4]
CDM	16264	47594	5.85	0.64	5.82	18	2.2087	0.1846	[31]
NSC	1461	2742	3.75	0.69	2.59	17	1.8486	0.4616	[33]

$$H = \frac{\langle \Gamma^2 \rangle}{\langle \Gamma \rangle^2} = \frac{\frac{1}{|V|}\sum_{x \in V} |\Gamma_x|^2}{\left(\frac{1}{|V|}\sum_{x \in V} |\Gamma_x|\right)^2}. \tag{8}$$

The assortativity coefficient, or assortative mixing, is a measure that assesses the preference of nodes in a network to attach to other similar ones [32]. The similarity definition can vary but it is usually computed using degree similarity between nodes as the correlation coefficient among the degrees of every pair of connected nodes. Considering the degree similarity, this score is computed as

$$\begin{aligned} r &= corr_{e_{x,y} \in E}(|\Gamma_x|, |\Gamma_y|) \\ &= \frac{\frac{1}{|E|}\sum_{e_{x,y} \in E} |\Gamma_x||\Gamma_y| - \left[\frac{1}{|E|}\sum_{e_{x,y} \in E} (|\Gamma_x| + |\Gamma_y|)/2\right]^2}{\frac{1}{|E|}\sum_{e_{x,y} \in E} (|\Gamma_x|^2 + |\Gamma_y|^2)/2 - \left[\frac{1}{|E|}\sum_{e_{x,y} \in E} (|\Gamma_x| + |\Gamma_y|)/2\right]^2}. \end{aligned} \tag{9}$$

A.2. Data sources

We have collected 22 networks from different sources and domains. These networks were carefully selected to cover a wide range of properties, including different sizes, average degrees, clustering coefficients, and heterogeneity indices. A summary of the collection of networks we used in our experiments can be found in Table 4. The collection of networks used in our experiments can be found at http://noesis.ikor.org/datasets/link-prediction. UPG is a power distribution network. HPD, YST, and CEG are biological networks. ERD, KNH, LDG, SMG, ZWL, HTC, CGS, CDM, NSC, and GRQ are co-authorship networks for different fields of study. HMT, FBK, and ADV are social networks. UAL is an airport traffic network. EML is a network of individuals who shared emails. PGP is an interaction network of users of the Pretty Good Privacy algorithm. BUP is a network of political blogs. Finally, INF is a network of face-to-face contacts in an exhibition.

A.3. Evaluating link prediction methods

A 5-fold cross-validation was carried out to measure the performance of link prediction techniques. Link prediction methods require a set of links to be used as a priori information to perform the inference process.

The original set of links E of each network was partitioned into k non-overlapping subsets $\{E_1, \ldots, E_k\}$ of equal size. On the ith iteration, only one of these sets was retained as validation set $E^V = E_i$ while the remaining sets were joined and used as training set $E^T = E - E_i$. Hence, $E^T \cup E^V = E$ and $E^T \cap E^V = \emptyset$. This process was repeated k times in order to use each subset once as validation set. The performance scores computed during each iteration were averaged to obtain a single score.

A missing link is formally defined as a link belonging to test set E^T. A non-observed link is defined as a link in the set $U_G - E^T$, where U_G is the complete graph of size $|V|$ containing the $\frac{|V||V-1|}{2}$ possible links that would exist if the network were fully connected. Finally, a non-existent link is defined as a link from the set $U_G - E$.

The results obtained by a supervised machine learning algorithm can be summarized using a confusion matrix. A confusion matrix shows all possible actual and predicted class combinations. In the context of link prediction, a true positive (TP) is a predicted link that actually belongs to the test set, a false positive (FP) is a predicted link that does not belong to the validation set, a false negative (FN) is a non-predicted link that belongs to the test set, and a true negative (TN) is a non-predicted link that does not belong to the validation set. These values are used to define different scores that measure different desirable properties in a link predictor.

We used precision as the performance measure for link predictors. Precision, also known as positive predictive value, is defined as the fraction of true positive links among the set of links predicted as true:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

Other measures have been used to evaluate the performance of link prediction techniques. For example, some authors have used the area under the curve (AUC) to compare different methods [26]. The AUC value is equal to the probability of the technique ranking a random missing link (a link from the validation set E^V) better than a random non-existent link (a link from $U_G - E$). Considering all pairs of missing and non-existent links, this value can be computed as

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

where n is the number of pairs, n' the number of pairs where the missing link was ranked better than the non-existent link and n'' the number of pairs where both links were ranked equally (for

example, by obtaining the same score or probability of existence). As expected, the AUC value for a random classifier must be around 0.5.

Recall has also been used by some authors [2]. Recall is similar to precision but considers the number of false negatives instead of the number of false positives. However, due to the way we evaluate link prediction techniques, it can be seen that FN increases as FP increases, so distinguishing between precision and recall would be unnecessary in our context (as is the use of alternative measures such as the *F*-score).

References

[1] L.A. Adamic, E. Adar, Friends and neighbors on the web, Soc. Netw. 25 (2003) 211–230.
[2] M. Al Hasan, V. Chaoji, S. Salem, M. Zaki, Link prediction using supervised learning, in: SDM'06: Workshop on Link Analysis, Counter-terrorism and Security, 2006.
[3] A.L. Barabási, R. Albert, Emergence of scaling in random networks., Science 286 (1999) 509–512.
[4] V. Batagelj, A. Mrvar, Pajek Datasets, 2006 http://vlado.fmf.uni-lj.si/pub/networks/data.
[5] 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 (2014) 750–764.
[6] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, et al., Topological structure analysis of the protein-protein interaction network in budding yeast, Nucleic Acids Res. 31 (2003) 2443–2450.
[7] A. Clauset, C. Moore, M.E. Newman, Hierarchical structure and the prediction of missing links in networks, Nature 453 (2008) 98–101.
[8] W. Cukierski, B. Hamner, B. Yang, Graph-based features for supervised link prediction, in: The 2011 International Joint Conference on Neural Networks (IJCNN) IEEE, 2011, pp. 1237–1244.
[9] N. Eggemann, S. Noble, The clustering coefficient of a scale-free random graph, Discrete Appl. Math. 159 (2011) 953–965.
[10] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, Y. Elovici, Link prediction in social networks using computationally efficient topological features, in: Privacy, Security, Risk and Trust (PASSAT), 2011 Third International Conference on Social Computing (socialcom), IEEE, 2011, pp. 73–80.
[11] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Stat. 11 (1940) 86–92.
[12] L. Getoor, C.P. Diehl, Link mining: a survey, ACM SIGKDD Explor. Newslett. 7 (2005) 3–12.
[13] N.Z. Gong, A. Talwalkar, L. Mackey, L. Huang, E.C.R. Shin, E. Stefanov, E.R. Shi, D. Song, Joint link prediction and attribute inference using a social-attribute network, ACM Trans. Intell. Syst. Technol. (TIST) 5 (2014) 27.
[14] R. Guimerà, M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks, Proc. Natl. Acad. Sci. U.S.A. 106 (2009) 22073–22078.
[15] Z. Huang, Link prediction based on graph topology: the predictive value of the generalized clustering coefficient, in: Workshop on Link Analysis (KDD), 2006.
[16] Z. Huang, X. Li, H. Chen, Link prediction approach to collaborative filtering, in: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, ACM, 2005, pp. 141–142.
[17] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.F. Pinton, W. Van den Broeck, What's in a crowd? Analysis of face-to-face behavioral networks, J. Theor. Biol. 271 (2011) 166–180.
[18] P. Jaccard, Étude comparative de la distribution florale dans une portion des alpes et des jura, Bull. Soc. Vaud. Sci. Nat. 37 (1901) 547–579.
[19] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, et al., Handling imbalanced datasets: a review, GESTS Int. Trans. Comput. Sci. Eng. 30 (2006) 25–36.
[20] V. Krebs, A Network of Books About Recent us Politics Sold by the Online Bookseller Amazon.com, 2008, Unpublished http://www.orgnet.com.
[21] J. Kunegis, Hamsterster full network dataset – KONECT., 2014 http://konect.uni-koblenz.de/networks/petster-hamster.
[22] E. Leicht, P. Holme, M.E. Newman, Vertex similarity in networks, Phys. Rev. E 73 (2006) 026120.
[23] J. Leskovec, J. Kleinberg, C. Faloutsos, Graph evolution: densification and shrinking diameters, ACM Trans. Knowl. Discov. Data (TKDD) 1 (2007) 2.
[24] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Am. Soc. Inf. Sci. Technol. 58 (2007) 1019–1031.
[25] W. Liu, L. Lü, Link prediction based on local random walk, Europhys. Lett.) 89 (2010) 58007.
[26] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A: Stat. Mech. Appl. 390 (2011) 1150–1170.
[27] V. Martí nez, C. Cano, A. Blanco, Prophnet: A generic prioritization method through propagation of information, BMC Bioinform. 15 (2014) S5.
[28] P. Massa, M. Salvetti, D. Tomasoni, Bowling alone and trust decline in social network sites, in: Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009. DASC'09, IEEE, 2009, pp. 658–663.
[29] J.J. McAuley, J. Leskovec, Learning to Discover Social Circles in Ego Networks, NIPS, 2012, pp. 548–556.
[30] M.E. Newman, Clustering and preferential attachment in growing networks, Phys. Rev. E 64 (2001) 025102.
[31] M.E. Newman, The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 404–409.
[32] M.E. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (2002) 208701.
[33] M.E. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (2006) 036104.
[34] M. Pavlov, R. Ichise, Finding experts by link prediction in co-authorship networks, FEWS 290 (2007) 42–55.
[35] S. Peri, J.D. Navarro, R. Amanchy, T.Z. Kristiansen, C.K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. Gandhi, M. Gronborg, et al., Development of human protein reference database as an initial platform for approaching systems biology in humans, Genome Res. 13 (2003) 2363–2371.
[36] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.L. Barabási, Hierarchical organization of modularity in metabolic networks., Science 297 (2002) 1551–1555.
[37] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.
[38] M.A. Serrano, M. Boguñá, R. Pastor-Satorras, A. Vespignani, Correlations in complex networks. Large scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Sciences, 2007, pp. 35–66.
[39] P. Colomer-de Simon, M. Boguná, Clustering of random scale-free networks, Phys. Rev. E 86 (2012) 026120.
[40] T. Sørensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons, Biol. Skr. 5 (1948) 1–34.
[41] S. Virinchi, P. Mitra, Similarity Measures for Link Prediction Using Power Law Degree Distribution, Neural Information Processing, Springer, 2013, pp. 257–264.
[42] C. Wang, V. Satuluri, S. Parthasarathy, Local probabilistic models for link prediction, in: Seventh IEEE International Conference on Data Mining, 2007. ICDM, IEEE, 2007, pp. 322–331.
[43] D.J. Watts, S.H. Strogatz, Collective dynamics of small-world'networks., Nature 393 (1998) 440–442.
[44] F. Wilcoxon, Individual comparisons by ranking methods, Biom. Bull. (1945) 80–83.
[45] T. Zhou, L. Lü, Y.C. Zhang, Predicting missing links via local information, Eur. Phys. J. B: Condens. Matter Complex Syst. 71 (2009) 623–630.

Víctor Martínez is a researcher at the Intelligent Databases and Information Systems research group in the Department of Computer Science and Artificial Intelligence at the University of Granada, Spain. He received a B.E. degree in Computer Engineering from the University of Granada in 2013 and a M.Sc. degree in Soft Computing and Artificial Intelligence from the University of Granada in 2014. His research interests include networks analysis, data mining and machine learning.

Fernando Berzal is an associate professor in the Department of Computer Science and Artificial Intelligence at the University of Granada. Previously, he had been a visiting research scientist at the data mining research group led by Jiawei Han at the University of Illinois at Urbana-Champaign. His research interests include model-driven software development, software design, and the application of data mining techniques to software engineering problems. He received his Ph.D. in Computer Science from the University of Granada in 2002 and he was awarded the Computer Science Studies National First Prize by the Spanish Ministry of Education in 2000. He is a senior member of the ACM and also a member of the IEEE Computer Society.

Juan-Carlos Cubero is a full professor in the Department of Computer Science and Artificial Intelligence at the University of Granada. He received his Ph.D. in Computer Science from the University of Granada in 1994. He has lectured in several European Universities and he has published several books in Computer Science and more than 30 papers in JCR journals. His research interests include database design, data mining, and software modeling. He has served as PC member in about 50 international conferences and he has actively participated in the organization of different conferences and workshops. He has also been the leader of a Spanish consortium participating in a European Eureka project.