mathematics

MDPI

Article

Discovering Correlation Indices for Link Prediction using Differential Evolution

Giulio Biondi [1,*,†] and Valentina Franzoni [2,*,†]

[1] Department of Mathematics and Computer Science, University of Florence, 50121 Florence, Italy
[2] Department of Mathematics and Computer Science, University of Perugia, 06123 Perugia, Italy
* Correspondence: giulio.biondi@unifi.it (G.B.); valentina.franzoni@dmi.unipg.it (V.F.)
† These authors contributed equally to this work.

Received: 11 October 2020; Accepted: 19 November 2020; Published: 23 November 2020

Abstract: Binary correlation indices are crucial for forecasting and modelling tasks in different areas of scientific research. The setting of sound binary correlations and similarity measures is a long and mostly empirical interactive process, in which researchers start from experimental correlations in one domain, which usually prove to be effective in other similar fields, and then progressively evaluate and modify those correlations to adapt their predictive power to the specific characteristics of the domain under examination. In the research of prediction of links on complex networks, it has been found that no single correlation index can always obtain excellent results, even in similar domains. The research of domain-specific correlation indices or the adaptation of known ones is therefore a problem of critical concern. This paper presents a solution to the problem of setting new binary correlation indices that achieve efficient performances on specific network domains. The proposed solution is based on Differential Evolution, evolving the coefficient vectors of meta-correlations, structures that describe classes of binary similarity indices and subsume the most known correlation indices for link prediction. Experiments show that the proposed evolutionary approach always results in improved performances, and in some cases significantly enhanced, compared to the best correlation indices available in the link prediction literature, effectively exploring the correlation space and exploiting its self-adaptability to the given domain to improve over generations.

Keywords: evolutionary algorithms; binary correlation; topological similarity; similarity of structure; evolutionary optimisation

1. Introduction

Link Prediction (LP) is a branch of Complex Networks science that aims at explaining the evolutionary dynamics of a network, looking at possible supplementary connections which can be established between entities (nodes) in the network. A common approach to LP is to introduce a definition of similarity between entities and to calculate similarity values accordingly, between all pairs of still non-connected nodes. In the ranking induced by the similarity rates, pairs ranked prime represent relationships with a higher formation likelihood. The image of a network at time t used to compute similarities is called *training network*, the information deriving from the ranking is tested on the *test network*, representing the status of the same network at a future time-step $t + 1$. The concept of similarity is central to the problem; in literature, various definitions are available, including semantic [1] and topological [2] similarity. The former evaluates similarity according to features of the nodes; intuitively, two nodes are as similar as their feature values are. The latter looks at the position of nodes in the network, either limiting the analysis to a k-depth bounded local neighborhood [3], or considering the whole network at once; e.g., [4,5] the broadly used Jaccard [6] and Adamic-Adar [7] indices. Important characteristics to consider for different approaches are the requirements, e.g., the number of training items for the learning phase; the possibility of reading and analyzing the process steps as well as the result, thus the readability of results; the result type, boolean, rank or absolute value with particular details. The choice of the approach will consider such requirements and adapt the technique and setting to the goal. In this work, we study the class of topological similarities, focusing on measures based on the shared local neighborhood (i.e., common neighbors), given that semantic similarity measures can also be mapped to topological ones [8], thus can be included in the same point of view. Similarities of depth 2, e.g., Resource Allocation and Adamic-Adar, have been demonstrated in the literature to be more effective in terms of prediction ability than other more straightforward measures [8]. However, this does not apply to all domains, and simple measures, e.g., Common Neighbours or Jaccard, often can outperform more elaborate ones. It looks like no all-purpose neighborhood-based similarity ratio, able to effectively capture the peculiar characteristics of each different domain, is available in the literature for a general application on every domain. Two research questions emerge:

1. How can the contribution of the best-performing indices in the literature on a given domain be exploited together?
2. Is it possible to modify indices to adapt them to any single domain, to reflect its specific link formation mechanisms?

To the best of our knowledge, the only attempt to answer the first research question is a plain linear combination of well-known indices [9], where the weights regulating the contribution of each index are evolved using the covariance matrix adaptation evolution strategy [10] for numerical optimization. This linear combination can be identified as a preliminary definition of a meta-correlation, but its adaptability power to different domains is limited. Our approach contributes to finding original meta-correlations evolving basic ones using Differential Evolution (DE) [11], where an added value is provided evolving the whole meta-correlation instead of using a plain linear combination of measures. Among the existing approaches to the problem of link prediction, we have chosen to build a meta-correlation based on the best indices in the literature and to adapt them to any domain using the Differential Evolution algorithm. DE is suitable for our goal for its readability and differentiation since our aim is finding a generalized meta-correlation metric to be applied to any domain without prerequisites of knowledge, density and connection of the graph. For example, methods based on full knowledge of the graph are very difficult to apply in large graphs, so the analysis of the nodes neighborhood is certainly more convenient [3]. Among these link prediction techniques, each of which can be better than others for different contexts, simple measurements often have decent results, but in the literature, many techniques are present for enhancing performances, with different variations. The Quasi-common proximity approach [3,12] varies the basic measurements in the graph to evaluate them at point 2 of the graph and is applicable to any topological similarity measure. Path-based heuristic approaches, such as the Heuristic Semantic Walk [13], calculate the similarity of potential nodes, applicable in link prediction, by choosing on the basis of semantic heuristics the direction for the graph navigation, adding partial randomization to avoid loops. Recently, some works combine topological and semantic similarity [1,8,14] to predict links in specific domains, e.g., co-authorship networks, providing techniques that could be applied also in other domains. Adapting the approach to many different similarity measures, very satisfying results are obtained, predicting links on the basis of sub-graphs [12,14] around nodes connected by each potential link, especially when semantic features are present but also using semantic measures mapped to the graph topology [1]. While topological and semantic approaches can exploit the characteristics of the network by recommending the network structure, on the other hand, approaches based on deep learning can be very performing, but require a very high number of training elements and provide results without any possibility for the researcher to analyze the process, which can be considered a black box. Some of these approaches, e.g., SEAL [5], use random sampling on potential links, not providing a complete list of rankings, to ease the computation. Unfortunately, all these techniques are not directly comparable, not only because of the different goals and approaches but because each of them uses proper evaluation metrics, which are different for each approach and not overlapping. The used similarity metrics vary based on the graph structure or features, and anyway, domain-specific characteristics do not allow a direct comparison where tests are made on different data sets. The choice of the right approach will vary in different contexts and goals, but it will be primarily based on the requirements of each approach. In real-world applications, e.g., where a company needs a correlation metric to exploit link prediction on any domain without the need for professional resources to set up different learning algorithms for each possible domain, it is useful to build meta-correlations with generalization capabilities.

The paper structure is the following: in Section 2 a formal definition of meta-correlation indices is given, the related state of the art is presented for the basic correlation metrics, and our proposed novel meta-correlations are presented in detail; Section 3 provides in-depth information for the experiment reproducibility and setting, including network preprocessing and partition, a description of the data sets where the experiments are exploited, and the setting of the Differential Evolution pipeline. Section 4 presents the experimental results and discussion; Section 5 concludes the paper.

2. Meta-Correlation Indices

Correlation indices have been defined by experts with different backgrounds to capture the peculiar properties of specific domains, and only afterward used in other application domains, e.g., biology, sociology and psychology. For example, the Dice (or Gleason, or Sørenson) index has been applied initially to ecological population data, while Simple Matching has been used to measure the level of agreement between two psychologists and Tanimoto in Chemoinformatics to analyze interaction fingerprints [15]; a large corpus of indices is available in the literature [16]. Regarding the domain of Link Prediction, various measures have been proposed and applied in previous works. Particular measures, e.g., Adamic-Adar [7] index, were purposefully developed for Link Prediction applications, while other ratios have been adapted to LP, e.g., the Jaccard [6] coefficient, initially used in biology and then in LP.

Formally, let x_1 and x_2 be two events or objects and F a set of features; most indices define the similarity between x_1 and x_2 as a function of four parameters a, b, c and d, which count the presence or absence of each $f \in F$. More specifically, $a(d)$ is the number of features available(not available) in both x_1 and x_2; b, and c counts the features occurring in, respectively, x_1 or x_2 only. Several indices can be seen as variations of a basic syntactic structure, where the input changes in terms of multiplicative coefficients and applied operators, e.g., summation and subtraction.

The framework introduced in this paper aims at optimizing the prediction strength of correlation indices by defining binary correlation meta-indices, which exploit structural similarity to create populations of correlation indices. The resulting indices are thus evolved using the Differential Evolution (DE) [17] algorithm. Binary correlation meta-indices are parametric formulas that subsume sets of correlation indices which include well-known indices for specific parameter values. Their parameters and structure fully characterize a meta-index; thus, a parameters assignment effectively defines a specific instance of the selected meta-index. Let for instance

$$\mu = \frac{\alpha a}{\beta a + \gamma b + \delta c + \epsilon}$$

where the meta-correlation index, μ, can subsume both the Sokal and Sneath-1 index when $\alpha = \beta = 1$, $\epsilon = 0$, and $\gamma = \delta = 2$, and the Common neighbors index when $\alpha = \epsilon = 1$, and $\beta = \gamma = \delta = 0$. Each possible assignment of values for the coefficients tuple of the meta-index represents then a valid and unique correlation index, while the meta-index itself represents a class of correlation indices composed of all the possible five-tuple values assignments for $\alpha, \beta, \gamma, \delta, \epsilon$. In the proposed framework, let μ meta-index used for Link Prediction on domain D, with n parameters $(c_1, \ldots, c_n)$:

- the population is composed a set of m vectors $v_1, \ldots, v_m$ of length n, each representing a correlation instance of the class subsumed by μ;

- the fitness function is any evaluation metric, e.g., precision, AUC, ROC, determining the capabilities of an individual for the Link Prediction task in the domain D.

One of the central focal points of our approach is that we designed two meta-correlations to subsume sets of well-known indices and incorporate them, combining the contribution of first-order and second-order features. The goal of the design of the experiment will then be to investigate whether evolving meta-correlation indices can adapt to the peculiar characteristics of a data set where they evolve.

Let (V, E) a network, where V is a set of nodes and E is the set of edges, $E \subseteq V \times V$, we define $\Gamma(u)$ where $u \in V$ as the set of neighbours of node u in the network G. Let u and v nodes of a network (V, E), the first-order features we considered are a=$|\Gamma(u) \cap \Gamma(v)|$, the number of Common Neighbours between u and v, b=$|\Gamma(u)| - |\Gamma(u) \cap \Gamma(v)|$ (resp. c=$|\Gamma(v)| - |\Gamma(u) \cap \Gamma(v)|$), the number of nodes connected only to u (resp. v) and $d = |V| - (a + b + c + 2)$, the number of the other nodes in the network, not connected to u nor to v. The second-order features, i.e., features that consider properties of nodes at distance 2 from u or v, are the well-known in the Link Prediction literature [7] Adamic-Adar similarity score

$$a_1 = \sum_{n \in \Gamma(u) \cap \Gamma(v)} \frac{1}{log|\Gamma(n)|} \tag{1}$$

the Pseudo-Adamic-Adar$_1$ score

$$b_1 = \sum_{n \in \Gamma(u) \setminus (\Gamma(u) \cap \Gamma(v))} \frac{1}{log|\Gamma(n)|}, \tag{2}$$

and the Pseudo-Adamic-Adar$_2$ score

$$c_1 = \sum_{n \in \Gamma(v) \setminus (\Gamma(u) \cap \Gamma(v))} \frac{1}{log|\Gamma(n)|}, \tag{3}$$

for neighbours connected respectively to u or v only.

Equations (4) and (5) show the two meta-indices formulas, while Tables 1 and 2 the subsumed indices and the corresponding parameter values.

$$similarity(u,v)_{\mu 1} = \frac{\alpha a + \beta b + \beta c + \gamma d + \delta a_1 + \epsilon b_1 + \epsilon c_1}{\zeta a + \eta b + \eta c + \theta d + \iota a_1 + \kappa b_1 + \kappa c_1 + \lambda 1} \tag{4}$$

$$similarity(u,v)_{\mu 2} = \frac{\alpha a + \beta a^2 + \gamma ab + \gamma ac + \delta bc + \epsilon a_1 + \zeta b_1 + \zeta c_1}{\eta a + \theta a^2 + \iota ab + \iota ac + \kappa bc + \lambda a_1 + \mu b_1 + \mu c_1 + \nu 1}. \tag{5}$$

Table 1. $\mu 1$ subsumed indices.

Index Name	Index Formulation	Coefficients
Jaccard	$\frac{a}{a+b+c}$	1,0,0,0,0,1,1,0,0,0,0
Dice	$\frac{2a}{2a+b+c}$	2,0,0,0,0,2,1,0,0,0,0
Sokal&Sneath-1	$\frac{a}{a+2b+2c}$	1,0,0,0,0,1,2,0,0,0,0
Sokal&Sneath-2	$\frac{2(a+d)}{2a+b+c+2d}$	2,0,2,0,0,2,1,2,0,0,0
Roger&Tanimoto	$\frac{a+d}{a+2(b+c)+d}$	1,0,1,0,0,1,2,1,0,0,0
Faith	$\frac{a+0.5d}{a+b+c+d}$	1,0,0.5,0,0,1,1,1,0,0,0
Sokal&Sneath-3	$\frac{a+d}{b+c}$	1,0,1,0,0,0,1,0,0,0,0
Kulczynski-1	$\frac{a}{b+c}$	1,0,0,0,0,0,1,0,0,0,0
Gower&Legendre	$\frac{a+d}{a+0.5(b+c)+d}$	1,0,1,0,0,1,0.5,1,0,0,0
Adamic-Adar	see Equation (1)	0,0,0,1,0,0,0,0,0,0,1

Table 2. $\mu 2$ subsumed indices.

Index Name	Index Formulation	Coefficients	Notes
Cosine	$\frac{a}{\sqrt{(a+b)(a+c)}}$	1,0,0,0,0,0,0,1,1,1,0,0,0	without square root
Sorgenfrei	$\frac{a^2}{(a+b)(a+c)}$	0,1,0,0,0,0,0,1,1,1,0,0,0	
Mountford	$\frac{a}{0.5(ab+ac)+bc}$	1,0,0,0,0,0,0,0,0.5,1,0,0,0	
McConnaughey	$\frac{a^2-bc}{(a+b)(a+c)}$	0,1,0,-1,0,0,0,1,1,1,0,0,0	
Johnson	$\frac{a}{a+b}+\frac{a}{a+c}$	0,2,1,0,0,0,0,1,1,1,0,0,0	
Kulczynski$_{\mathrm{II}}$	$\frac{a^2+0.5ab+0.5ac}{a^2+ab+ac+bc}$	0,1,0.5,0,0,0,0,1,1,1,0,0,0	
Adamic-Adar	see Equation (1)	0,0,0,1,0,0,0,0,0,0,0,0,1	

3. Experiments

The goal of the experiment is to investigate whether evolving meta-correlations can adapt them to the peculiar characteristics of the particular domain where they evolve.

3.1. Network Preprocessing

The usual approach for Link Prediction experiments is to divide a data set into two parts, which are conventionally defined training set and test set; the test set, usually amounting to 10–20% of the data set, is used to evaluate the performance of models built using the knowledge provided in the training set. For this work, the followed approach is instead to split the data set into three parts. First, the data set is split in training and validation set E_{TR+V} and test set E_{TE}, following a 90:10 ratio; then, k folds are generated from E_{TR+V}, building k (E_{TR}, E_V) pairs, which will be used to evolve different correlations each. Before this phase, the networks are pre-processed to remove elements such as self-loops and isolated nodes, since both do not add any contribution to similarity scores calculated using local neighborhood-based measures. Directed networks are transformed into undirected networks: when there is a connection in at least one way between two nodes, they are connected in the pre-processed network.

3.2. Data Sets

The framework has been tested on four data sets, widely used in link prediction literature. The data sets represent two main domains with some diversity in each data set. The first domain comprises CA-GrQC [18] and Netscience [19], representing the co-authorship domain with two diverse networks, respectively including papers published in the General Relativity and Quantum Cosmology categories between 1993 and 2004, and in the area of Network Science. The other two data sets, ia-radoslaw-email [20] and email-eu-core [18,21] are two e-mail exchange networks, thus representing a digital communication domain, the first between employees of a European institution, and the other of a medium-sized company. Such domains have been considered representative of authors' social networks (i.e., co-authorship) and communication networks, to test the proposed approach on real similar domains. For each domain, two instances have been chosen for sharing some similarities, to show how the metrics used to create the meta-correlation do not have themselves similar results even in similar networks, and to test if our evolved meta-correlation indices can better forecast the link creation both in similar (i.e., about the same domain), and in diverse networks (i.e., about different domains).

3.3. Settings for Differential Evolution

The population members for Differential Evolution (DE) are evolved using the information available in the training set, and their fitness is calculated on the validation set. Precision, i.e., the proportion of properly ranked edges among the top-k edges, is used as a fitness metric, while k is set to $|E_V|$; a perfect predictor would rank all the positive edges as first. Edges in the test set are not available during the evolution process, effectively appearing as non-existent. The number of generations G was experimentally set to 300, as it was observed that further iterations did not provide any improvement. The mutant weighting factor F and the Crossover constant CR have been set, respectively, to 0.9 and 0.5, according to literature [11]. The core part of the population P is composed of instances of correlation indices p_i which subsume known indices; additional individuals are obtained by applying random noise $n, -0.25 < n < +0.25$ to the coefficients of such correlations to explore more extensively the correlations space.

An observed problem that could arise using Differential Evolution is the loss of diversity when there is a total or too high consensus on one parameter value; this could happen for the proposed meta-correlation instances because the parameters frequently present the same values for subsumed indices. Introducing noise-altered population members allows overcoming this problem. The population for μ_1 amounts to 27 individuals, of which nine represent known indices and the rest two variations each; for μ_2, six known correlations are considered, along with three variations each. Two DE variants have been tested in this work, namely RAND/1/EXP and RAND/1/BIN according to the conventional DE naming scheme. Both employ a random selection for the individuals used in the mutation phase, and use one pair of individuals (hence RAND/1); EXP and BIN refer to the adopted crossover schemes, meaning respectively, exponential and binary [22].

3.4. Algorithm

After pre-processing the network, the evolutionary phase to derive new correlation indices begins. For each fold f, the best individual of the population springing at the end of the DE execution is compared versus known correlation indices. The combined knowledge available in E_{TR} and $+E_V$ is used as ground truth to rank probable edges. Since E_{TE} was excluded from the training process, we can test the performance on E_{TE}, to assess the potential of the correlation in predicting future edges. The framework structure pipeline for Differential Evolution is shown in Algorithm 1.

Algorithm 1: Framework structure for Differential Evolution (DE).

```
Pre-process the network;
Initialize the population of meta-correlation instances;
for $f \leftarrow 1$ to $K$ do
    for $g \leftarrow 1$ to $G$ do
        for $p \in P$ do
            $y_i \leftarrow$ generate_offspring$(p_i)$;
            Rank potential edges according to $y_i$ using information in $E_{TR}$;
            Evaluate the fitness $f(y_i)$ on $E_V$;
            if $f(y_i) > f(p_i)$ then
                Replace $p_i$ with $y_i$;
        Save the best individual $p_b$;
Test $p_b$ on $E_{TE}$ using combined information from $E_{TR}$ and $E_V$;
```

4. Experiments Results

In this section, we present the results of the experiments in terms of Precision (see Section 3.3). Other suitable metrics include AUC [23], and SRD [24,25]. In Tables 3 and 4, the precision values of the best individuals across all the folds, evolved following strategies RAND/1/EXP and RAND/1/BIN respectively, are compared to known correlation indices, namely Common Neighbours (CN), Jaccard and Adamic-Adar (AA), for each data set.

Table 3. Precision of best individual, RAND/1/EXP strategy. Best performance in bold.

Data Set	CN	Jaccard	Adamic-Adar	μ_1	μ_2
Netscience	0.425455	0.501818	0.654545	**0.661818**	0.654545
CA-GrQc	0.369220	0.366460	0.489303	**0.554865**	0.492754
ia-radoslaw-email	0.412308	0.273846	0.418462	0.436923	**0.470769**
email-eu-core	0.195395	0.191039	0.221531	**0.276914**	0.27318

Table 4. Precision of best individual, RAND/1/BIN strategy. Best performance in bold.

Data Set	CN	Jaccard	Adamic-Adar	μ_1	μ_2
Netscience	0.425455	0.501818	0.654545	**0.676364**	0.647273
CA-GrQc	0.369220	0.366460	0.489303	**0.554865**	0.491373
ia-radoslaw-email	0.412308	0.273846	0.418462	0.427692	**0.480000**
email-eu-core	0.195395	0.191039	0.221531	**0.276291**	0.274424

The best improvements in performance are obtained on the CA-GrQc dataset, on which the best individual for μ_1 performs noticeably better than the reference measures. Similar behaviour can be observed on the email-eu-core data set, where both μ_1 and μ_2 achieve higher scores than the best performing index, AA. Slight improvements are also noticeable on the netscience data set, where μ_1 ranks first. Differently from the other data sets, on ia-radoslaw-email the best performing meta-correlation is $\mu 2$, demonstrating sensible improvements, while μ_1 yields performance comparable to other measures. For all the data sets, the precision values are higher than the reference measures, for μ_1, μ_2, or both.

In Tables 5 and 6, the average precision and variance of the best individuals for each fold on $E_{TR} + E_V$ for all the data sets are reported. Although the discovered correlations in some cases greatly differ in terms of their coefficients values, all of them achieve better performances, both for μ_1 and μ_2. This probably hints at a correlation space with many separated local maxima with similar values.

Table 5. Average precision and Variance, RAND/1/EXP strategy. Best performance in bold.

Data Set	Average Precision μ_1	Variance Precision μ_1	Average Precision μ_2	Variance Precision μ_2
Netscience	**0.618182**	0.001404	0.596000	0.001207
CA-GrQc	**0.531401**	0.000217	0.470393	0.000163
ia-radoslaw-email	0.404923	0.000436	**0.432308**	0.002176
email-eu-core	0.263535	0.000111	**0.263908**	0.00005

Table 6. Average precision and Variance, RAND/1/BIN strategy. Best performance in bold.

Data Set	Average Precision μ_1	Variance Precision μ_1	Average Precision μ_2	Variance Precision μ_2
Netscience	**0.628727**	0.001313	0.591273	0.000503
CA-GrQc	**0.532850**	0.000125	0.475362	0.000240
ia-radoslaw-email	**0.405538**	0.000344	0.398154	0.008970
email-eu-core	**0.271624**	0.000014	0.260547	0.000047

The intuition about local maxima is reinforced by looking at graphs in Figures 1 and 2 where the dynamics of the evolutionary process are illustrated for the Netscience and ia-radoslaw-email data sets, for both meta-correlations. Charts on the left show the performance improvement from generation 1 to 300 for meta-correlation μ_1, on the right for μ_2. Each line represents the evolution on a fold, following the DE RAND/1/EXP strategy; for readability, only the behavior on a subset of folds is shown.

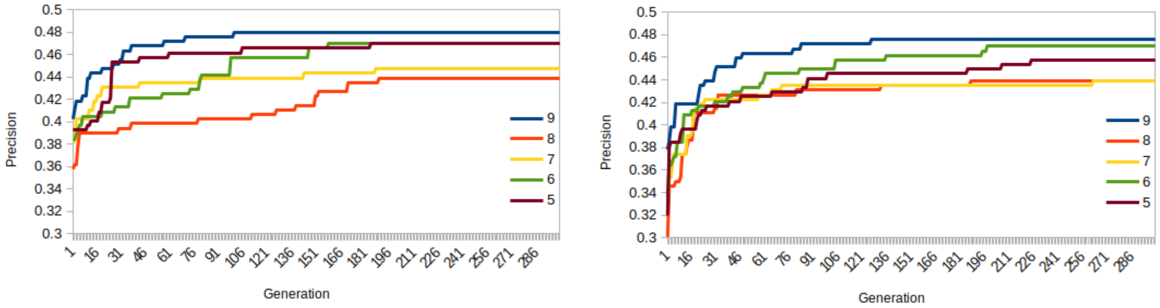

Figure 1. Precision of best individual over generations on Netscience data set. On the X axis the generation g, and on the Y axis the precision value of the best individual in the population at generation g. Each line shows the evolution dynamics on a specific fold.

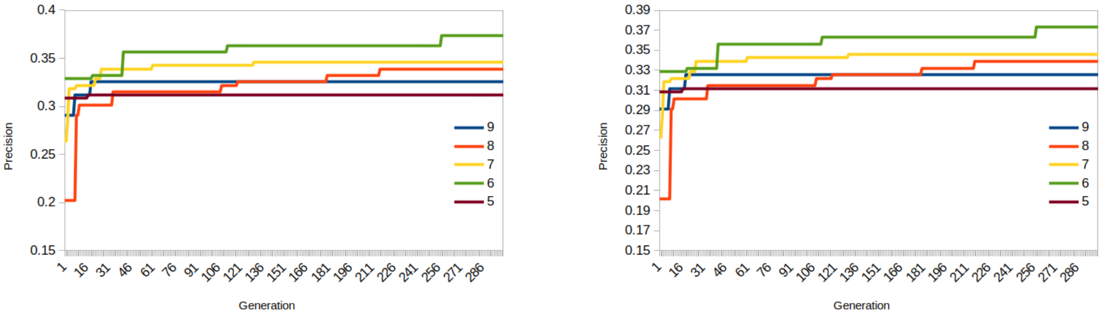

Figure 2. Precision of best individual over generations on ia-radoslaw-email data set.

It can be noticed that the correlations space presents plateaus on which the fitness function returns the same score: this behaviour slows down the evolution process. Other combinations of data sets, correlations and strategies report in the literature similar behaviour [11]. In the vast majority of cases, the performance improvements from the first to last generation noticeable, ranging from 3% to more than 10%. In few cases the system shows a less substantial increment in precision, e.g., in fold 5 for the ia-radoslaw-email data set, where the improvement is visible but unimportant; this can be due to the k-fold validation split.The evolution of meta-correlation is always enhanced with respect to the single measures on every data set. Our approach is not directly comparable with other techniques of Link Prediction, which are experimented in the literature on other data sets, using different similarity measures and other evaluation metrics, e.g., the Quasi-common neighbors [3] or the SEAL heuristic-learning technique [5], to cite other techniques possibly capable to be adapted to include both topological and semantic similarity, because the main focus of our work is not to obtain a better link prediction on a single domain, as happens with the cited approaches, but to discover a meta-correlation (which can be also used in other techniques like Quasi-common neighborhood) with the ability to adapt evolving, to every domain. Even if the comparison of a particular measure may eventually perform better on a particular domain, that measure does not have the power to adapt its performance to other domains, even similar ones. Our meta-correlation, based on the link prediction power of any desired measure, can evolve it to its best performance on general domains, which is our main goal. A general comparison can be done on the knowledge requirements: the more graph knowledge is required (i.e., the broader the analyzed common-neighborhood), the stronger computational capabilities are required for the prediction. Adaptability to the contexts of all domains is the main enhancement of our proposal.

5. Conclusions

In this work, we presented a framework based on evolutionary algorithms for Link Prediction. Differential Evolution (DE) is used to evolve the coefficients of parametric meta-correlations formulas to design domain-centered indices. Meta-correlations identify new classes of correlations; each component is identified by a different meta-correlation parameter vector, also subsuming well-known indices for specific parameters assignment. During the DE evolution process of the population of meta-correlation parameter vectors, new correlation indices are discovered, with prediction capabilities tailored to a specific data set, i.e., environment. Experiments show that the system can integrate the contribution of different features to discover new correlation indices that improve the precision value when compared to link prediction indices existing in the literature. The initial research questions now have answers: with our method, it is possible to have a general meta-correlation striking a good balance between being adaptive (more than standard indices), and less computationally intensive (e.g., than other learning-based methods achieving similar or better performance with particular metrics on particular domains), exploiting the contribution of the best-performing indices adapted to the specific link formation mechanism of each possible domain.

Future works aim at extending the experiment domain to assess the extra capabilities of the system in various research domains where discovery and optimization of correlation indices are an explicitly crucial point. To obtain the best results, it would be needed to re-implement all the approaches in the literature for link prediction or at least those we mentioned in the introduction, using the same similarity measures, the same evaluation metrics and the same data sets, to have a complete and realistic direct survey comparison among them, independently of the objective and the application context of the developer and the user.

Author Contributions: Conceptualization, G.B. and V.F.; methodology, G.B. and V.F.; software development, G.B.; validation, G.B. and V.F.; formal analysis, G.B. and V.F.; investigation, G.B. and V.F.; data curation, G.B. and V.F.; writing—original draft preparation, G.B. and V.F.; writing—review and editing, V.F.; visualization, G.B.; supervision, V.F.; project administration, G.B. and V.F.; funding acquisition,G.B. and V.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by the Italian Ministry of Research under PRIN Project “PHRAME” Grant n.20178XXKFY.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Franzoni, V.; Milani, A. Structural and semantic proximity in information networks. *Lect. Notes Comput. Sci.* **2017**, *10404*, 651–666.
2. Liben-Nowell, D.; Kleinberg, J. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.* **2007**, *58*, 1019–1031. [CrossRef]
3. Chiancone, A.; Franzoni, V.; Niyogi, R.; Milani, A. Improving Link Ranking Quality by Quasi-Common Neighbourhood. In Proceedings of the 15th International Conference on Computational Science and Its Applications (ICCSA 2015), Banff, AB, Canada, 22–25 June 2015; pp. 21–26. [CrossRef]
4. Martínez, V.; Berzal, F.; Cubero, J.C. A Survey of Link Prediction in Complex Networks. *Acm Comput. Surv.* **2016**, *49*, 69. [CrossRef]
5. Zhang, M.; Chen, Y. Link Prediction Based on Graph Neural Networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31, pp. 5165–5175.
6. Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Del Soc. Vaudoise Des Sci. Nat.* **1901**, *37*, 547–579.
7. Adamic, L.A.; Adar, E. Friends and neighbors on the Web. *Soc. Netw.* **2003**, *25*, 211–230. [CrossRef]
8. Biondi, G.; Franzoni, V. Semantic Similarity Measures for Topological Link Prediction. In *Computational Science and Its Applications—ICCSA 2020*; Springer International Publishing: Cham, Switzerland, 2020; pp. 132–142. [CrossRef]
9. Bliss, C.A.; Frank, M.R.; Danforth, C.M.; Dodds, P.S. An evolutionary algorithm approach to link prediction in dynamic social networks. *J. Comput. Sci.* **2014**. [CrossRef]
10. Hansen, N.; Ostermeier, A. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In Proceedings of the IEEE International Conference on Evolutionary Computation, Nagoya, Japan, 20–22 May 1996; pp. 312–317. [CrossRef]

11. Biondi, G.; Milani, A.; Baia, A.E. Differential Evolution of Correlation Indexes for Link Prediction. In Proceedings of the 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 12–14 December 2018; pp. 1483–1486.
12. Chiancone, A.; Franzoni, V.; Li, Y.; Markov, K.; Milani, A. Leveraging zero tail in neighbourhood for link prediction. In Proceedings of the 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Singapore, 6–9 December 2015; pp. 135–139. [CrossRef]
13. Franzoni, V.; Milani, A. Heuristic semantic walk for concept chaining in collaborative networks. *Int. J. Web Inf. Syst.* **2014**, *10*, 85–103. [CrossRef]
14. Franzoni, V.; Lepri, M.; Li, Y.; Milani, A. Efficient Graph-Based Author Disambiguation by Topological Similarity in DBLP. In Proceedings of the 2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Laguna Hills, CA, USA, 26–28 September 2018; pp. 239–243. [CrossRef]
15. Rácz, A.; Bajusz, D.; Héberger, K. Life beyond the Tanimoto coefficient: Similarity measures for interaction fingerprints. *J. Cheminf.* **2018**, *10*, 48. [CrossRef] [PubMed]
16. Seung-Seok, C.; Sung-Hyuk, C.; Tappert, C.C. A Survey of Binary Similarity and Distance Measures. *J. Syst. Cybern. Infor.* **2010**, *8*, 43–48.
17. Storn, R.; Price, K. Differential Evolution—A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *J. Glob. Optim.* **1997**. [CrossRef]
18. Leskovec, J.; Kleinberg, J.; Faloutsos, C. Graph evolution: Densification and Shrinking Diameters. *Acm Trans. Knowl. Discov. Data* **2007**, *1*, 2-es. [CrossRef]
19. Newman, M.E. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. Stat. Nonlinear Soft Matter Phys.* **2006**, *74*. [CrossRef] [PubMed]
20. Michalski, R.; Palus, S.; Kazienko, P. Matching Organizational Structure and Social Network Extracted from Email Communication. In *Lecture Notes in Business Information Processing*; Springer: Berlin/Heidelberg, Germany, 2011; Volume 87, pp. 197–206.
21. Yin, H.; Benson, A.R.; Leskovec, J.; Gleich, D.F. Local higher-order graph clustering. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017. [CrossRef]
22. Lin, C.; Qing, A.; Feng, Q. A comparative study of crossover in differential evolution. *J. Heuristics* **2011**, *17*, 675–703. [CrossRef]
23. Chen, B.; Hua, Y.; Yuan, Y.; Jin, Y. Link Prediction on Directed Networks Based on AUC Optimization. *IEEE Access* **2018**, *6*, 28122–28136. [CrossRef]
24. Héberger, K. Sum of ranking differences compares methods or models fairly. *TrAC Trends Anal. Chem.* **2010**, *29*, 101–109. [CrossRef]
25. Kollar-Hunek, K.; Heberger, K. Method and model comparison by sum of ranking differences in cases of repeated observations (ties). *Chemom. Intell. Lab. Syst.* **2013**, *124*, 139–28136. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.