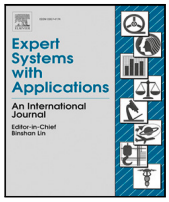
Expert Systems With Applications 255 (2024) 124417



Contents lists available at ScienceDirect

Expert Systems With Applications

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





Analysis and prediction of the Horizon 2020 R&D&I collaboration network

Zsolt T. Kosztyán [a,*], Ferenc Király [a], Attila I. Katona [a], Tibor Csizmadia [b], Beáta Fehérvölgyi [c]

[a] Department of Quantitative Methods, Egyetem str. 10, Veszprém, H-8200, Hungary
[b] Department of Management, Egyetem str. 10, Veszprém, H-8200, Hungary
[c] Department of Innovation Management, Egyetem str. 10, Veszprém, H-8200, Hungary

ARTICLE INFO

Keywords:
Link prediction
R&D&I projects
Collaboration network analysis
Network-based database

ABSTRACT

This study presents a novel link prediction model for the analysis and prediction of collaboration networks in the context of H2020 projects. This model incorporates various databases and expands the list of variables used in the analysis. In this study, the accuracy and interpretability of both generic and nongeneric link prediction algorithms are examined through a comprehensive analysis. Additionally, collaborative groups that exhibit stronger linkages than initially anticipated are examined, and their prevalence across different regions is evaluated. In this study, it was found that the collaboration network within H2020 has significant concentration and fragmentation levels. Moreover, several key factors that strongly influence collaboration, including prior and concurrent project involvement, earned contribution, and the economic and technological characteristics of participating businesses, were identified. This study additionally demonstrated that cooperative communities exhibit geographical concentration in western Europe, whereas nongeneric models offer improved predictive accuracy for identifying relationships. This research contributes to the field of cooperation network analysis and offers guidance to decision-makers on strategies to promote future collaborations in the context of FP.

1. Introduction

Global research networks are emerging due to the international division of R&D&I processes through, among other processes, international and interorganizational collaboration. A collaboration network is a process in which organizations have a broader capacity to access ideas, knowledge, and solutions from a variety of sources (Sachwald, 2008). The open innovation approach stresses that external collaboration increases the innovative capacity of organizations (Chesbrough, 2003), particularly by exploiting external sources of creativity and knowledge (Enkel, Gassmann, & Chesbrough, 2009) and learning from knowledge transfer (Secundo, Toma, Schiuma, & Passiante, 2019). The Horizon 2020 program also stresses prioritizing international and interorganizational collaboration (Novitzky et al., 2020; Veugelers et al., 2015). The importance of international collaboration has been further recognized by the European Union (EU) and many other countries in the global response to the COVID-19 pandemic (Cai, 2023).

A collaboration network is applicable to a wide variety of application areas. These networks have been used in many domains and disciplines, such as bibliography and library science (Elmagarmid, Ipeirotis, & Verykios, 2006; Malin, Airoldi, & Carley, 2005), bioinformatics (Airoldi, Blei, Fienberg, Xing, & Jaakkola, 2006), e-commerce (Liu & Kou, 2007), Industry 4.0 (Camarinha-Matos, Fornasiero, & Afsarmanesh, 2017), the internet and web science (Adafre & de Rijke, 2005; Zhu, Hong, & Hughes, 2002), and manufacturing (Camarinha-Matos, Afsarmanesh, Galeano, & Molina, 2009; MacCarthy & Jayarathne, 2012). In many of the above applications, the networks are not necessarily social networks; rather, they can be the internet, information networks, or biological entity networks.

In addition, collaboration networks have been applied at different levels. In studies on countries, patent databases (De Prato & Nepelski, 2014), publication and citation databases (Guan, Zuo, Chen and Yam, 2016; Guns & Wang, 2017), and international trade databases (Guan, An, Gao, Huang and Li, 2016; Liu, Dong, Ding, Wang, & Zhang, 2020) have been used, while for studies on regions, publication databases (Chuanming, Yutian, Xiaoli, & Lu, 2017) and patent databases (De Noni, Orsi, & Belussi, 2018) have been used. In a research paper on cities, a publication database (Guns & Rousseau, 2014) was used. Finally, in studies focusing on the organizational level, publication databases (Chang & Huang, 2013; Han et al., 2014; Huang, Zhu, Zhang, Zhou, & Jia, 2018; Lande et al., 2020; Lee, Seo, Choe, & Kim, 2012) and patent databases (Chen, Qu, & Chi, 2021) have mostly been used. In these studies, one database has primarily been used; at the same time, in some papers, databases such as PATSTAT and ORBIS (Cséfalvay &

* Corresponding author.
E-mail addresses: kosztyan.zsolt@gtk.uni-pannon.hu (Z.T. Kosztyán), kiraly.ferenc@gtk.uni-pannon.hu (F. Király), katona.attila@gtk.uni-pannon.hu (A.I. Katona), csizmadia.tibor@gtk.uni-pannon.hu (T. Csizmadia), fehervolgyi.beata@gtk.uni-pannon.hu (B. Fehérvölgyi).

https://doi.org/10.1016/j.eswa.2024.124417
Received 11 March 2024; Received in revised form 7 May 2024; Accepted 3 June 2024
Available online 26 June 2024
0957-4174/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Gkotsis, 2022; Mahnken & Moehrle, 2018; Tarasconi & Menon, 2017) and ORBIS and CORDIS have been combined. To our knowledge, there has been no combination of corporate data (ORBIS), collaboration data (CORDIS), patent data (PATSTAT), or economic background data (EUROSTAT) at the regional or organizational level. Nevertheless, without the combination of the aforementioned data, accurate collaboration prediction is not possible.

Link prediction is widely used to find missing links in social networks, and based on existing connections, predictions are made on whether pairs of unconnected nodes will be connected in the future (Chen et al., 2021). In addition, social network link prediction has applications in other domains, such as bioinformatics, information retrieval and e-commerce (Hasan & Zaki, 2011). Existing link prediction methods are commonly grouped into three categories: similarity-based methods, probability-based methods, and learning-based methods (Chen et al., 2021). Similarity-based methods measure the similarity of unconnected pairs of nodes to predict whether a link should consist of local and global information (Liu et al., 2016; Sun, Hu, Yang, Yao, & Yang, 2017; Xu & Yin, 2017). Probability-based methods assess the possibility of nonexistent links by constructing statistical probability models (Chi, Yin, Dong, & Dong, 2019; Yang, Dong, & Xie, 2014). Learning-based methods, such as machine learning (Li, Wang, Jiang, Zhang, & Yan, 2018; Mohan, Venkatesan, & Pramod, 2017; Zhang, Zhang, Yuan, & Zhang, 2016) and metaheuristics (Bastami, Mahabadi, & Taghizadeh, 2019), use intelligent frameworks and algorithms for link prediction. Since these methods can predict the links between unconnected nodes, they are applicable for partner selection in patent collaboration networks (Chen et al., 2021). Hasan and Zaki (2011) emphasized that there are many techniques for link prediction, ranging from feature-based classification and kernel-based methods to matrix factorization and probabilistic graphical models. These techniques differ from each other in terms of model complexity, prediction performance, scalability, and generalizability.

Link prediction has an important role in scientific collaboration networks, and several studies have aimed to predict collaboration between partners, such as authors or organizations. Chuanming et al. (2017) established a collaboration recommendation model based on scientific collaboration network neighbors and paths. They conducted an empirical investigation to test the model at the individual, institutional and regional levels. In 2019, Wang, Han, Sun, and Pan (2019) built eight weighted algorithms to predict potential scientific collaborations. They found that weighted prediction algorithms based on fusion indicators yield better prediction results. In this article, they proposed and combined two indicators with four commonly used indicators, the institutional document frequency (IDF) and the institutional cumulative cooperation ratio (ICCR), to denote the similarity of the authors' references in choosing cooperating institutions. In these articles, the analyses were performed at the organizational level. Later, Lande et al. (2020) used an algorithm to predict collaborations between scientists. The presented method is based on a heterogeneous information network model. This model consists of authors and keywords from articles extracted from the Web of Science database. Chen et al. (2021) tested eight different link prediction approaches to explore partner selection in China's interorganizational patent cooperation network. Qi et al. (2022) dealt with choosing the right collaboration partner for innovation. They created a framework based on topic analysis and link prediction. In their study, they highlighted three subcategories of literature-based methods, one of which is network analysis/link prediction methods using single or multiple network indicators or in combination with other external feature indicators based on bibliographic data. In this study, they used a link prediction method with a fusion network of author and patentee organizations to predict potential collaborations. The other two subcategories are the index methods based on bibliographic information and methods that use natural language processing (NLP) techniques to analyze the "intrinsic" content of the literature. There are recent deep learning ling prediction methods, such as graph neural networks (GNNs). GNNs have gained popularity for link prediction in complex networks by effectively capturing the structural information of the network. They can learn node embeddings and predict links based on the graph topology and the patterns inferred from the existing connections. This allows GNNs to make accurate predictions about potential links between nodes in the network, even when only partial information is available (Zhou et al., 2020). Knowledge graph embeddings (KGE) can embed entities and relations in knowledge graphs into low-dimensional vector spaces, enabling link prediction and knowledge discovery tasks to be performed (Ge, Wang, Wang, Kuo, et al., 2024). Path-based algorithms leverage the semantic paths between entities in knowledge graphs to predict missing links and infer new relationships between entities (Rossi, Barbosa, Firmani, Matinata, & Merialdo, 2021). Graph convolutional networks (GCNs) extend convolutional neural networks to graph-structured data, enabling the modeling of node relationships for link prediction tasks in various domains such as social networks, citation networks, biological networks, recommender systems, and security (Zhang, Tong, Xu, & Maciejewski, 2019). GCNs have shown promising results in link prediction by effectively capturing the local and global structural information of the graph (Zhou et al., 2020). Moreover, graph attention networks and graph recurrent networks have also emerged as variants of GNNs that have demonstrated outstanding performance on various deep learning tasks, including link prediction.

Although the aforementioned link prediction studies achieved considerable results in collaboration analysis, the problem related to prediction accuracy is still an outstanding shortcoming in this research field (Wang, Xu, Wu, & Zhou, 2015). There are two reasons for unsatisfactory prediction accuracy: (1) model-based and (2) dataset-based. From the model side, the computational complexity causes crucial problems. While GNN, KGE and GCN have shown promising results in link prediction, the scalability of these models is problematic. As the size of the network increases, the computational complexity of training and inference with GNNs becomes a bottleneck. Additionally, the effectiveness of GNNs in capturing long-range dependencies and patterns in complex networks remains a topic of ongoing research. Since the investigated collaboration network contained 20,172 organizations, 237,084 edges, and 44 features, in line with the findings of Wang et al. (2015), GNNs did not seem to be an effective method for addressing this problem. In addition to the scalability problem, analysis and interpretation of feature importance are not supported well by neural networks; however, this is also an important aim in this context. These findings are in line with the results of Chen et al. (2021) since the prediction precision achieved by numerous algorithms fluctuated around approximately 0.75. The lack of an accurate link prediction model inhibits strategic planning at both the organizational and consortium levels. From the database side, a lack of comprehensiveness can lead to inaccurate prediction results. To date, no study has analyzed combined databases, including the ORBIS, CORDIS, PATSTAT and EUROSTAT datasets. This leads us to the problem of low prediction accuracy since important variables and relationships may be missed due to the use of limited databases. The unsatisfactory prediction accuracy makes it challenging to make reliable predictions related to EU FP collaborations.

Moreover, in addition to the accuracy of the prediction models, there is an additional gap in this research field. The available studies have focused on the overall prediction performance; however, they have not focused on the detection of outliers (significantly weaker or stronger connections compared to the prediction) across the network. If outliers are not analyzed, information is not available about the specific regions/actors; therefore, customized support is not feasible in EU FPs. This makes it difficult for the EU to improve the efficiency of collaboration.

To bridge the aforementioned research gaps, the following research question is posed in this paper:

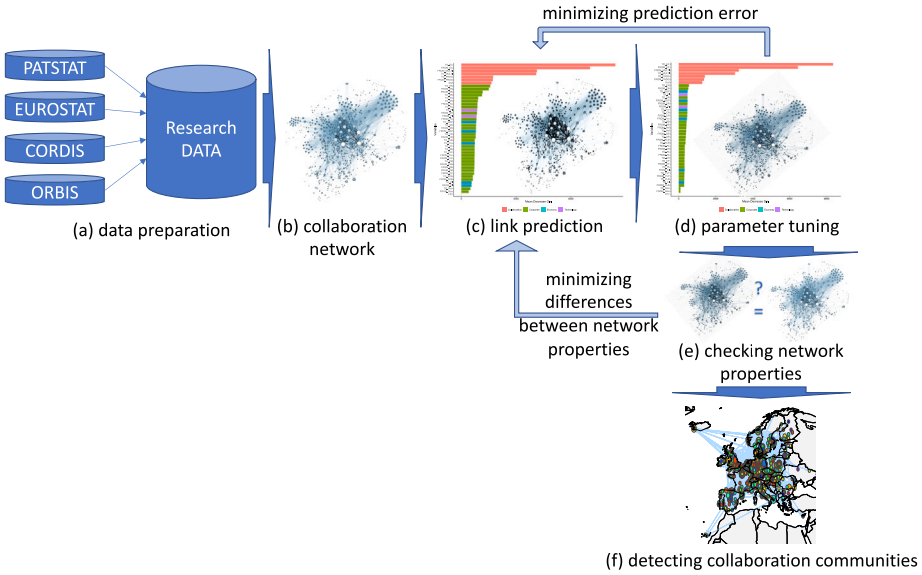


Fig. 1. Simplified flow chart of the proposed prediction method.

RQ Is it possible to model with high accuracy and explain the collaboration and communities between organizations in FPs using a comprehensive dataset?

To answer this research question, this paper proposes a new model to analyze and predict the collaboration network of H2020 by combining multiple databases and extending the analyzed variable set with new factors. Building an accurate link prediction model is beneficial for decision-makers in predicting future consortia across the FP. Determining the key influential factors allows decision-makers to support the formation of future FP collaborations. Finally, this study also identifies outlier collaboration communities (with stronger or weaker links than predicted), which are relevant for decision-makers in controlling collaboration formation across the FP. The aforementioned elements are in line with EU directions, which emphasize the strengthening of collaboration among actors at the country and organization levels (Novitzky et al., 2020).

Our contribution to the literature is threefold:

C_1 We have created an accurate model for predicting links in collaboration, including indications derived from corporate, collaboration, patent and economic databases.

C_2 We identified the key economic, technological, and collaborative factors that influence collaboration.

C_3 The proposed economic null model is able to specify the outlier collaboration communities where the collaboration is stronger than expected based on the prediction.

This research presents an accurate model for examining and predicting collaboration within the Horizon 2020 framework. This is achieved by combining various databases and expanding the range of variables with multiproject and program membership characteristics (C_1). The study identifies the most influential factors associated with collaboration, including the extent of involvement in previous partnerships, indicators of completion structure, and corporate indicators (C_2). This study revealed that specific communities had more outlier collaboration patterns than expected by the model (C_3). Overall, these findings can support decision-makers in effectively managing and enhancing cooperation creation within EU FPs.

The research question was addressed as follows. The research method is introduced in Section 2. Section 3 reports the results of the study. Section 4 discusses the results of the study, and finally, Section 5 provides the summary and conclusions.

2. Materials and methods

Fig. 1 shows the main steps of the link prediction process.

In the first step, a research database is prepared, and four datasets are combined (see Fig. 1(a)). The PATSTAT dataset contains the patents used as a proxy indicator for describing the technology background, and the EUROSTAT dataset is used to describe the economic background of a company. The CORDIS dataset is used to consider collaboration indicators, and the ORBIS dataset is applied to consider a company's performance (see details in Section 2.1). Based on the prepared research data, the collaboration network is specified (see Fig. 1(b)), where the nodes represent organizations and the edges between the nodes represent the earned projects. Based on the research data, technology, economic, corporate, and collaboration attributes are assigned to each node; the geographic distance and the former presence of collaboration are assigned to each edge, which is used in link prediction (see Fig. 1(c)). Link prediction methods rank collaboration, technology, corporate, and economic indicators according to their importance for prediction (see Section 2.2). Several generic and nongeneric (black-box) link prediction methods (see Section 2.2) are compared to find the best link predictions. Nongeneric methods are tuned to minimize the prediction error (see Fig. 1(d)). All the link prediction methods are validated when the network structure indicators between the original and predicted collaboration networks are compared (see Fig. 1(e)). After these loops, the tuned parameters of the machine learning methods, sample sizes, and best link prediction method are provided. Link prediction methods already consider collaboration, corporate, technology, and economic indicators. In addition to link prediction, the specified model can be used to specify *collaboration communities*. Within these communities, the collaborations are even stronger than what the prediction model predicts (see Fig. 1(f)).

2.1. Data

In this study, we used three different kinds of data sources. The main data source was the CORDIS Horizon 2020 dataset,[1] which contains information on more than 25 thousand earned projects and more than 150 thousand participants. In this database, 86% of participants were European participants, as other countries also participated in projects.

[1] https://cordis.europa.eu/.

Table 1
Applied indicators.

	v	Indicator*	Description	Data source	Level	Type
Node dataset	m_1	TA	Total assets	ORBIS	Org.	Corporate
	m_2	SR	Solvency ratio (Asset based) (%)	ORBIS	Org.	
	m_3	SH	Shareholders' funds	ORBIS	Org.	
	m_4	RB	ROE using P/L before tax (%)	ORBIS	Org.	
	m_5	RCB	ROCE using P/L before tax (%)	ORBIS	Org.	
	m_6	PM	Profit margin (%)	ORBIS	Org.	
	m_7	PLF	P/L for period	ORBIS	Org.	
	m_8	PLF	P/L before tax	ORBIS	Org.	
	m_9	OR	Operating revenue	ORBIS	Org.	
	m_{10}	FAs	Fixed assets	ORBIS	Org.	
	m_{11}	EN	Number of employees	ORBIS	Org.	
	m_{12}	CR	Current ratio	ORBIS	Org.	
	m_{13}	CF	Cash flow	ORBIS	Org.	
	m_{14}	Size	Size of the company	ORBIS	Org.	
	m_{15}	GDP	GDP/capita in purchasing power priority	EUROSTAT	NUTS3	Economy
	m_{16}	CI	Corruption Index. Scale: 0 (highly corrupt) to 100 (very clean).	EUROSTAT	NUTS1	
	m_{17}	Pov	Poverty	EUROSTAT	NUTS2	
	m_{18}	PI	Patents	PATSTAT	NUTS3	Technology
	m_{19}	EC	Earned Contribution	CORDIS	Org.	Collaboration
	m_{20}	MULTI	Mean of the multiproject membership value	CORDIS*	Org.	
	m_{21}	PROG	Mean of the program membership value	CORDIS*	Org.	
Edges	i	FROM	The ID of the partner company	CORDIS/ORBIS	Org.	
	j	TO	The ID of the coordinator	CORDIS/ORBIS	Org.	
	D	Dist	Geographical distance between two organizations	CORDIS/ORBIS	Org.	
	FP7		Any earned projects in a collaboration between companies	CORDIS	Org.	

In addition, 72% of participants are organizations, and 28% are individuals. The database contains the start/end times of the project, a description of the project goals, and organizational data. The organizational data are connected to the ORBIS[2] dataset, which contains financial data from more than 24 million companies in Europe. However, this database does not contain data on sole traders or micro companies. The entity analysis between the CORDIS and ORBIS datasets was performed by the owner of the ORBIS dataset, Bureau van Dijk Company, based on company names, addresses, and geocodes. From the EUROSTAT databases, county-level economic data (nomenclature of territorial units for statistics (NUTS3)), such as gross domestic product (GDP)/capita (NUTS3 level), were also considered. The total number of patents was obtained from the PATSTAT database. After these databases were combined, 20,172 organizations were identified (Table 4). The simplified database schema can be found at Figure 5 in Supplementary material A.

The specified network dataset contains the following vertex and edge datasets (see Table 1). There are four variable categories. (1) *Collaboration* indicators include multiproject (MULTI) and program membership values, the distances between organizations (DIST), and a dummy variable (FP7) for whether these organizations collaborated in the former 7th Framework Programme (FP7) program. (2) *Corporate* indicators include the main profit & loss and balance sheet records of the companies. (3) *Economy* background indicators include GDP/capita, the corruption index (CI), and poverty (Pov) for the country/region. (4) The number of patents (PI) is a proxy indicator of the *technology* indicator.

Table 1 shows that most variables are organizational-level variables from the employed data sources. Only the GDP per capita (m_{15}), which describes the economic environment, and the number of patents (m_{16}), which describes the technological environment, are county-level (NUTS3) data. Except for two indicators (MULTI and PROG), the remaining indicators can be determined directly from the source databases. The MULTI and PROG indicators must be calculated based on Kosztyán, Katona et al. (2022)'s study (see the details in Section 2.2.1). The multiproject memberships are related to the number of earned projects running in parallel, where the organizations of the consortium are (at least partly) common (see Eq. (1)). The program membership value relates to how a project builds up the former earned project outcomes (see Eq. (2)). These membership values can be assigned to a project; however, the mean value of organized projects can be calculated for each organization. Suppose the mean value of a multiproject membership is high for a company. In that case, the company organizes several projects that are running in parallel simultaneously. In contrast, a high mean value of program membership indicates that the organization has experience with previous related success projects.

We used the CORDIS database, which includes project and company information, as the main data source. We added the BvdId field to uniquely identify companies and link them to the ORBIS database, which contains economic and geographical data. We also used the PATSTAT database to obtain the number of patents by NUTS3 region and the EUROSTAT database to obtain the GDP, GDP per capita, corruption index, and poverty information for the NUTS region. We integrated these data into one database using NUTS codes as the common key.

Our final dataset consists of 20,172 nodes and 237,084 links. The geocodes are based on the address of the registered office and site and were obtained from the open street map application. The database of vertices contains 201 variables and 59,761 observations. If an organization occurs in more than one project, the repetition is aggregated by BvD_ID. The columns of vertices are summarized below. The names of the applied indicators are listed in Table 1.

NodeNo | BvD_ID | projectID | startDate | endDate | CompamyName | NationalLegalForm | City | Region | Country | AdditionalAddress | NUTS1 | NUTS2 | NUTS3 | Size_class | ProjectRcn | role | CordisID | endOfParticipation | ecContribution | street | postCode | organizationURL | vatNumber | contactForm | ContactType | ⋯ | contactFaxNumber | membership_multi | membership_program | BvDmatchquality | TA2010... TA2019 | ⋯ | CF2019 | PI_NUM2011... PI_NUM2019 | CI_NUM2011... CI_NUM2019 | GDPNUTS3_2011... GDPNUTS3_2019 | POVNUTS2_2011... POVNUTS2_2019 | Lat | Lon

We created two data files for the network edges, one with direction and one without. The direction was from the participant to the coordinator of the project. We also added a field to indicate whether the entities had previously participated in FP7 projects. The database of edges contains 5 variables and 237,054 observations. The columns of the edge database are summarized below.

[2] https://www.bvdinfo.com.

BvD_ID_From | BvD_ID_To | ProjectId | IsDirected | FP7connection

When organizations collaborate on more than one project in this network, multiple arcs are created. These edges need to be simplified to predict links. Simultaneously, this knowledge is not entirely eradicated, as it is encoded in the quantity of parallel projects organized by the consortium (MULTI) and the extent to which they rely on prior outcomes (PROG) during the summits.

From the proposed database, we created one training sample and one test sample. Importantly, in the case of databases specified for link prediction purposes, the database will most often be unbalanced because two organizations are far more often not in contact with each other than they are involved in joint projects. Most machine learning methods will produce biased results if this problem is not addressed. For this reason, we first selected pairs of organizations with collaborative relationships in the test and training databases, where the splitting ratios were 0.3 (test) and 0.7 (training). Then, we randomly selected p pairs of organizations that did not have a collaborative relationship in the database. Finally, the parameterized model was tested on the full database. We chose the p number with the best accuracy and F1 scores for $N = 100$ runs for most methods for the entire project. After tuning parameter p, the best F1 scores were achieved by setting $p = 1.302585$. When no links are assumed between organizations, the accuracy (Accuracy Null) is 0.767704. Thus, we expect all link prediction methods to have significantly higher accuracy than this value.

To employ link prediction, the applied node indicators (see Table 1) are linked to both the head (e.g., TO_TA.., TO_PROG) and the tail (e.g., FROM_TA.., FROM_PROG) of the edges. Therefore, only one table is applied for the link prediction, where the first column is a logical value for whether the nodes are connected or not. It follows the node indicators for the source and the tail. Last, the geographic distances and the logic variable of the former collaboration in FP7 are stated at the end of the dataset.

2.2. Methods

2.2.1. Calculation coefficients from the precedence structure

Distance matrices are calculated among the projects based on the overlap duration (d_t), precedence (d_p), ownership (d_o), volume (d_v), and description (d_{xy}). For the detailed calculation of the distance metrics, see Kosztyán, Katona et al. (2022). By using the pairwise distance values, the multiproject and program membership values can be calculated:

$$\mathcal{M}_m(p_i) = \max_j \left\{ \left(1 - d_t(p_i, p_j)\right) \cdot \left(1 - d_o(p_i, p_j)\right) \cdot \left(1 - d_v(p_i, p_j)\right) \right\} \quad (1)$$

$$\mathcal{M}_p(p_i) = \max_j \left\{ \left(1 - d_p(p_i, p_j)\right) \cdot \left(1 - d_{xy}(p_i, p_j)\right) \right\} \quad (2)$$

where p_i and p_j denote the ith and jth projects within the H2020 framework, respectively, and $\mathcal{M}_m(p_i)$ and $\mathcal{M}_p(p_i)$ are the multiproject and program membership values related to the p_i project, respectively.

2.2.2. Applied machine learning methods for link prediction

Accuracy measures help us assess the performance of a classification model. A confusion matrix compares a model's predictions with the actual results.

We calculated the following accuracy measures: *accuracy* (the proportion of correct predictions to the total number of cases), *precision* (the proportion of correct positive predictions to the total number of predicted positives), *sensitivity* (the proportion of correct positive predictions to the total number of actual positives), *specificity* (the proportion of correct negative predictions to the total number of actual negatives), *prevalence* (the proportion of positive) cases in the population, and *F1 score*, the harmonic mean of precision and sensitivity.

These measures are commonly used to evaluate classification models. The most important measure depends on the problem and the goals of the analysis. In our study, we attempted to predict collaboration between organizations. We focused on the accuracy and F1 score, but we also reported all the other measures.

Link prediction problem. $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denotes the collaboration graph between organizations, where $\mathcal{V}$ denotes the set of nodes (i.e., organizations) and $\mathcal{E}$ denotes the set of directed edges from the partner organization to the coordinator. In this study, each edge is a binary edge. If $e_{i,j} \in \mathcal{E}$ and $v_i, v_j \in \mathcal{V}$, organization i is a partner of organization j in at least one common project.

All employed machine learning problems must estimate the following link prediction problem:

$$e_{i,j} \sim \hat{y}_{i,j} = f(x_{i,j}) \quad (3)$$

where $x_{i,j} = [m_{i,k}, m_{j,k}, D_{i,j}, FP7_{i,j}]$ is a feature vector. $\hat{y}_{i,j}$ is the predicted link between node i and node j. f is the prediction function, $m_{i,k}$ is the value of indicator $k \in 1, 2, \ldots, 19$ of organization i, and $D_{i,j}$ is the geographical distance between two organizations.

$$FP7_{i,j} = \begin{cases} 1 & \text{if organization } i \text{ and organization } j \text{ collaborated in FP7} \\ 0 & \text{otherwise} \end{cases}.$$

Generic machine learning methods. Binary LogR is a widely used statistical technique in machine learning for binary classification problems that models the relationship between a set of input features and a binary target variable and produces a probability estimate of the target variable being positive (e.g., 1) given the input features.

In link prediction problems in collaboration networks, binary LogR can be used to predict whether a link (i.e., a collaboration) will occur between two nodes (i.e., organizations) based on their attributes ($m_1 - m_{19}$) and the structure (i.e., $D_{i,j}$) of the network. The target variable is binary, indicating whether a link exists between two nodes.

Binary LogR can be a powerful tool for link prediction in collaboration networks because it can capture complex nonlinear relationships between the input features and the target variable and can handle high-dimensional feature spaces. This technique also provides a probabilistic estimate of the likelihood of a link between two nodes, which can be useful in ranking candidate links by their probability of occurrence.

However, as in any machine learning technique, binary LogR has limitations and assumptions. For example, it is assumed that the relationship between the input features and the target variable is linear on the log-odds scale. This binary LogR may not perform well when the relationship is highly nonlinear or when the feature space is sparsely populated.

Discriminant analysis (DA), such as Binary LDA, is a statistical technique that is commonly used in machine learning for binary classification problems. A linear combination of features that can best discriminate between two classes is sought.

Binary LDA works by first estimating the mean and variance of each input feature for each of the two classes. Then, the linear discriminant function that maximizes the separation between the two classes is calculated. The linear discriminant function is defined as a linear combination of the input features, such that the distance between the means of the two classes is maximized and the within-class variance is minimized. The decision boundary between the two classes is then determined based on the value of the linear discriminant function. In the context of link prediction in a collaboration network between organizations, binary LDA can be used to identify which features of the organizations and the network are most important in predicting whether a link will exist.

Binary LDA has several advantages over other machine learning techniques in the context of link prediction in collaboration networks,

including its ability to handle high-dimensional feature spaces, simplicity and interpretability, and ability to capture the underlying structure of the data (Gu, Li, & Han, 2011). However, as in any machine learning technique, binary LDA has limitations and assumptions. LDA assumes that the relationship between the independent variables and the dependent variable is linear. DA also assumes that the independent variables are normally distributed within each group. If this assumption is violated, the results may be biased and inaccurate. DA assumes that the variances of the independent variables are equal across groups. If this assumption is violated, the results may be biased and inaccurate. In addition, DA can be sensitive to multicollinearity, where there is a high correlation between independent variables. This can lead to unstable estimates of the discriminant function.

Both linear regression and linear discriminant analysis attempt to estimate Eq. (4), which is a linear version of Eq. (3)

$$e_{i,j} \sim \beta_0 + \beta_D D_{i,j} + \beta_{FP7} FP7_{i,j} + \sum_{k:=1}^{19} \beta_{i,k} m_{i,k} \beta_{j,k} m_{j,k} \tag{4}$$

Instead of LDA, a QDA can be used to obtain better results for link prediction (Tharwat, 2016). The main benefit of using QDA instead of LDA is that QDA can handle nonlinear relationships between the features and the target variable (Tharwat, 2016).

Nongeneric machine learning methods. We used Support vector machine (SVM) (Hearst, Dumais, Osuna, Platt, & Scholkopf, 1998) to predict collaboration between organizations (Yu, Xu, & Wang, 2020). SVM is a discriminative algorithm that learns a decision boundary to separate data into classes. However, SVM does not output class membership probabilities. We used the feature vector x_{ij} for each pair of organizations $e_{i,j}$ and the SVM model to predict the label $\hat{y}$ for the pair. If $\hat{y} = 1$, SVM predicts a high likelihood of collaboration, and if $\hat{y} = -1$, SVM predicts a low likelihood of collaboration. The main hyperparameters of SVM are as follows Gold and Sollich (2003): *C*: The tradeoff between margin and error. *Kernel type*: This function transforms the data into a higher-dimensional space. Common kernels include linear, polynomial, Radial basis function (RBF), and sigmoid kernels. *Kernel coefficient*: This parameter controls the shape of some kernels. *Degree*. The degree of the polynomial kernel. *Gamma*. The width of the Gaussian kernel for the RBF kernel. *Class weights*: The weights to account for class imbalance. Moreover, SVM is powerful but has several limitations. For example, these methods do not scale well to large datasets, depend on the choice of kernel function, are sensitive to noisy data, and produce a binary classifier that is difficult to interpret.

We also used RF (Pal, 2005), which creates an ensemble of decision trees trained on random subsets of data and features. Each tree makes a prediction, and the final prediction is based on the majority vote. RF can handle nonlinear and high-dimensional data and provide information about feature importance.

The prediction function of RF for a pair of organizations ($e_{i,j} \sim \hat{y}_{i,j}$) can be formulated as follows:

$$e_{i,j} \sim \hat{y}_{i,j} = f(x_{i,j}) = mode([T_1(x_{i,j}), T_2(x_{i,j}), \ldots, T_M(x_{i,j})]) \tag{5}$$

where f is the prediction function, $x_{i,j}$ is the feature vector for the pair of organizations, and $T_1, T_2, \ldots, T_M$ are the decision trees in the ensemble.

The main hyperparameters of RF are Probst, Wright, and Boulesteix (2019): *n_estimators*: The number of trees in the forest. More trees can improve the performance but increase the cost. *max_depth*: The maximum depth of each tree. *min_samples_split*: The minimum number of samples needed to split a node. *min_samples_leaf*: The minimum number of samples at a leaf node. *max_features*: The number of features to consider for the best split. Setting this term to "sqrt" or "$\log_2$" can reduce the risk of overfitting by randomly selecting features. *bootstrap*: Whether to use bootstrapping to sample the data for each tree. Setting this hyperparameter to "true" means that each tree is trained on a random subset of the data, which can reduce the correlation between the trees.

We used LogR to identify the significant variables for link prediction. However, LogR assumes a linear combination between variables and cannot handle nonlinear relationships or nonseparable classes (Tonkin, Woodhams, Bull, Bond, & Santtila, 2012). We also used Boruta, a feature selection algorithm (Kursa, Jankowski, & Rudnicki, 2010) that compares the importance of each feature with that of shadow features. The shadow features are created by shuffling the values of each feature, which breaks the relationship with the target variable. Boruta uses a random forest classifier to determine the importance of each feature and removes the unimportant features, helping to prevent overfitting and improve the accuracy. RF is an ensemble learning method (Sagi & Rokach, 2018) that creates a forest of decision trees trained on random subsets of data and features. Each tree makes a prediction, and the final prediction is based on the majority vote. RF has several limitations (Ahmad, Basheri, Iqbal, & Rahim, 2018). For example, RF is a complex black-box model that is difficult to interpret. In addition, this model is computationally expensive and memory intensive, especially for large datasets or many trees. Therefore, there is a balance between bias and variance. Finally, this model may struggle to handle highly correlated features or imbalanced datasets.

Extreme gradient boosting (XGBoost) (Sheridan, Wang, Liaw, Ma, & Gifford, 2016) is a machine learning algorithm that uses decision trees as its base learners. These trees are trained in a sequence, where each new tree corrects the errors of the previous ones. This algorithm adds a regularization term to prevent overfitting and uses gradient boosting to update the weights. Several hyperparameters can be tuned to optimize the performance. XGBoost also has several limitations (Demir & Şahin, 2022). For example, overfitting can easily occur if the hyperparameters are not tuned well. Many hyperparameters need to be selected carefully. Moreover, this approach has limited handling of categorical variables and may require one-hot encoding. This approach can be computationally intensive for large datasets and requires careful feature engineering. In addition, XGBoost is a black-box model that is difficult to interpret (see Table 2).

Tuning of machine learning methods. We can tune the hyperparameters of machine learning methods to optimize their performance. Common methods for hyperparameter tuning are the *grid search*, *random search*, and *Bayesian optimization* methods (Alibrahim & Ludwig, 2021). A grid search evaluates all possible combinations of hyperparameter values using a metric and cross-validation. The best combination is selected as the optimal set of hyperparameters. A grid search can explore the hyperparameter space well, but it can be costly when there are many hyperparameters or large ranges of values. A random search randomly selects hyperparameters from a range or distribution, trains and evaluates a model using each combination. This process is repeated for several iterations or until a good result is achieved. A random search can often find optimal or near-optimal values faster than a grid search, but it is not guaranteed that the best combination will be found. Bayesian optimization builds a probabilistic model of the objective function (the metric to optimize) and uses it to guide the selection of hyperparameters. The model is updated with the results of each iteration, and exploration and exploitation are balanced. Bayesian optimization can efficiently search complex hyperparameter spaces by using probabilistic models. In this study, we used Bayesian optimization.

We followed Shahriari, Swersky, Wang, Adams, and De Freitas (2015) review in the setting of Bayesian optimization in the context of hyperparameter tuning. During initialization, the process started with an initial set of hyperparameters (see Table 3). Model fitting: A probabilistic surrogate model (in this case, the Gaussian process) is fitted to the objective function using the evaluated hyperparameters and corresponding performance metrics. An acquisition function (in this case, the expected improvement) is used to determine the next

Table 2
Advantages and limitations of machine learning techniques for link prediction.

Method	Advantages	Limitations
LogR	Simple and interpretable	Limited to linear decision boundaries
LDA	Handles multiple classes well	Assumes normality and equal covariances
SVM	Handles high-dimensional data well	Can be sensitive to choice of kernel function
RF	Handles nonlinear relationships well	Can overfit on noisy data
XGBoost	High accuracy and generalizability	Requires more data preprocessing and tuning

Table 3
The tuned hyperparameters of the machine learning algorithms and the search ranges applied during tuning.

Algorithm	Hyperparameter	Values
SVM	Classification machine	C, μ, One
	Kernel	Sigmoid, radial
	Kernel parameters (γ)	$2^{-5}, 2^{-4}, \ldots, 2^{5}$
	Cost of constraints violations	$2^{-5}, 2^{-4}, \ldots, 2^{5}$
RF	Number of estimators	20, 30, 40, 50, 75, 100, 120
	Split criterion	Gini, entropy
	Max depth of trees	10, 12, 14, 16, 18, 20, 25
	Minimum number of samples required to split an internal node	2, 3, 4
	Minimum number of samples required to be at a leaf node	1, 2, 3, 4
XGBoost	Number of estimators	20, 30, 40, 50, 75, 100, 120
	Max depth of trees	10, 12, 14, 16, 18, 20, 25
	Learning rate (η)	[0.01,0.5]
	Subsample ratio of columns when constructing trees	0.25, 0.5, 0.75, 1
	Subsample ratio of the training instance	0.25, 0.5, 0.75, 1
	L2 regularization (λ)	[−1,1]

set of hyperparameters to evaluate. The acquisition function balances exploration (sampling in uncertain regions) and exploitation (sampling in regions with high expected improvement). The selected hyperparameters are evaluated via cross-validation to obtain the performance metric. The new data points (hyperparameters and performance metric) are used to update the surrogate model, refining the probabilistic representation of the objective function. We repeat the previous two steps iteratively for a predefined number of iterations or until a convergence criterion is met. The process aims to find the set of hyperparameters that optimize the performance metric of the machine learning model.

2.2.3. Analyzed network properties

The benefit of calculating and comparing network indicators for the original and predicted collaboration networks is threefold. First, the comparison of network structure indicators highlights the extent to which link prediction methods have also succeeded in estimating the structure of the predicted network. Second, for non-black-box methods, we obtain a model for the prediction of collaborations and for the derived network indicators (Kosztyán, Király and Kurbucz, 2022). Third, in most cases, a so-called configuration model (Girvan & Newman, 2002) is used to form communities. It is assumed that collaborative network organizations are randomly connected. In addition, communities form where connections are denser between organizations than between individual communities (modules) (Newman & Girvan, 2004). However, if an edge prediction model is used as the null model (Gadar, Kosztyan, & Abonyi, 2018), communities can form where connections are even denser than those predicted by the model (Kosztyán et al., 2021).

To verify the prediction, we used both node-level (such as centrality measures) and network-level or structural indicators (such as centralizations, assortativities and modularity values) to compare the original and the predicted structures. The comparison of node-level indicators indicates not only the changes in the links between the predicted and the original networks but also the changes in embeddedness. The comparison of network-level indicators indicates structural changes.

The following node-level indicators are calculated to verify the prediction of the link. *DC*: The number of direct connections of a node divided by the maximum possible number of direct connections. This indicator identifies influential nodes that have many collaborators. *Betweenness centrality (BC)*: The number of shortest paths that pass through a node is divided by the total number of shortest paths. The nodes that act as intermediaries in the network are identified. *Eigenvector centrality (EVC)*: The sum of the centrality scores of the neighbors of a node. The nodes that are connected to other important nodes are identified. *Centrality*: The relative importance or influence of a node or an edge based on its degree, position, or relationship to other nodes or edges.

We also employed network-level indicators, such as *modularity*, which is the difference between the actual and expected number of links within communities. Groups of nodes that collaborate more with each other than with other nodes are identified. *Assortativity*: The correlation between the degrees of the linked nodes. The collaborative patterns between nodes with similar attributes are identified. Further details and formulas can be found in the Supplementary Materials B. *Centralizations*: Centralization in collaboration networks generally refers to the extent to which collaborations are concentrated around a few organizations.

The absolute value of the differences between the original and the predicted collaboration network's assortativity is also calculated for each prediction technique.

3. Results

3.1. Descriptive statistics

Table 4 shows the descriptive statistics of the collaboration network of the Horizon 2020 framework program.

The number of vertices (nodes) in the network is 20,172. These nodes represent the entities (at least small organizations) involved in the Horizon 2020 R&D&I projects, which have a European BvdID from the ORBIS database. There are 237,084 edges (links) in the network, representing collaboration in a project. The high number of components and relatively low density indicate that the network is not fully connected and that there are many small subnetworks. Moreover, the relatively low average distance between vertices suggests that the entities in the network are closely connected to each other in a single component. The relatively high average distance between entities when considering their geographical location suggests that the collaborations may be more international in nature. A high γ value suggests that the degree distribution of the network follows a power-law distribution,

Table 4
Descriptive statistics of network properties.

Property	Values
Vertices	20,172
Edges	237,084
Number of components	4363
Average distance (vertices)	3.37862
Average distance (km)	977.65710
Density	0.00045
γ	2.56377
Assortativity	−0.05226
Degree of centralization	0.12146
Betweenness centralization	0.10415
Eigenvector centralization	0.99259
Modularity	0.61364

Table 5
Main indicators in different company sizes.

Indicators\Size	Small	Medium	Large	Very large
TA	7180	31,834	38,512	3,009,231
OR	4782	29,133	11,258	1,789,507
PLB	399	650	1038	126,959
MULTI	0.0707	0.0872	0.0666	0.0961
PROG	0.4390	0.4450	0.4210	0.4410
N	6647	3528	7038	2959

meaning that there are relatively few entities with very high degrees (i.e., highly connected) and many entities with low degrees. A high gamma value indicates a high concentration. A negative assortativity value suggests that high-degree entities tend to be connected to low-degree entities rather than to other high-degree entities. A low negative assortativity value means that the tendency for organizations to collaborate with dissimilar organizations is weak. In other words, organizations in the network tend to collaborate with organizations that are similar to them in terms of the measured characteristic. However, the degree of similarity is not strong enough to produce a positive assortativity value. The relatively low degree of centrality suggests that the network is not strongly centralized around high-degree entities. A high eigenvector centrality value suggests that there are some entities that are particularly important in the network based on their connections to other important entities. The relatively low BC value suggests that there are no entities that are particularly important in controlling the flow of information between other entities. A high modularity value suggests that the network is strongly modular and that there are several distinct communities of closely connected entities in the network.

Table 5 shows several indicators and company sizes.

Table 5 shows that the resources of companies, such as total assets (TA), operational revenue (OR), and profit and loss values before tax (PLB), greatly differ, while the average multiproject (MULTI) and program (PROG) membership values are very similar. Very large and medium companies run more projects in parallel, but these values are relatively low in all cases. Nevertheless, a higher number of program membership values indicates that all kinds of companies like to be involved in a project. with a company that has previous experience.

3.2. Results of generic machine learning techniques

Table 6 shows standardized group means of linked (1) and unlinked (2) organizations (nodes) and the coefficients of the discrimination functions by partners and coordinators. Table 6 is separated into vertex-related (i.e., partner and coordinator) and edge-related (e.g., Dist and FP7) indicators. The variables are classified into four variable groups: corporate (c), economic (E), technological (T), and collaboration (C) variables.

The results of the corporate variables show that companies that establish collaborations with each other have better conditions, such as greater Total assets (TA) (see, e.g., both FROM_TA, TO_TA group means), Fixed assets (FAs), better Shareholder funds (SH), ROE using P/L before tax (%) (RB), ROCE using P/L before tax (%) (RCB), greater Profit margin (%) (PM), Operation revenue (OR), P/L before tax (PLF), and P/L for period (PLF). Only two corporate indicators are lower in the case of collaborative companies: Current ratio (CR) and Solvency ratio (asset based) (%) (SR). In addition, in the case of coordinators, the difference in group means is greater, which indicates that better corporate conditions are more critical when coordinators are involved.

Better conditions apply not only to company resources but also to the economic background. Collaborating companies may have a better economic background, such as higher GDP, lower poverty, or a cleaner economy (higher CI). Regarding the higher value of patents In collaborative organizations, reflecting the technological background indicates that the technological background also plays a vital role in collaboration.

On one hand, collaboration initiators can be linked to the structural characteristics of the network. After all, as we have seen, the network is a scale-free network in which the earned contribution and granted projects are concentrated in a few organizations, as indicated by the fact that other organizations prefer to be associated with those organizations where Earned contribution (EC) is higher. On the other hand, organizations that already run several supported projects (MULTI) and have previous successful project experience (PROG) prefer to cooperate.

Collaboration indicators for edges show that collaboration is less likely to occur between partners who are geographically further away. Moreover, if there has already been cooperation between organizations in the previous framework program, these organizations exhibit a greater willingness to cooperate. After splitting the training and test sets of the links, the accuracy on the test set was 0.77, and the F1 score was 0.85. This result is quite good for a generic algorithm. The QDA leads to better accuracy, which was 0.82, and the F1 score was 0.89. The results suggest that there are nonlinear connections between variables. However, LDA provides acceptable accuracy and describes the nature of the independent variables.

Table 6 also provides information on the significance of the explanatory variables. On both the partner (FROM) and coordinator (TO) sides, the corporate indicators RCB, PLF, PLF, FAs, Number of employees (EN), CR, and Cash flow (CF) were found to be insignificant . For the group of economic indicators, the corruption index was significant, and except for prior collaboration, all the collaboration indicators were significant. Because of the different ranges of the explanatory and explained variables, the odds ratio was close to 1. Therefore, the group mean values of LDA were more interpretable than those of LDA, and because of the small β-s, some coefficients changed signs compared to those of LDA. However, three of the significant variables yielded remarkable interpretations. In line with the results of LDA, it was found that larger organizations have a greater chance of collaborating. In addition, LogR indicated that a larger company has a 19.38% greater chance of being a partner and a 20.54% greater chance of becoming a coordinator in a collaboration. The parallel completions indicator (MULTI) provided the best odds ratio. It was also found that an organization that is already involved in a collaborative project is 17 times more likely to be a partner in a new collaboration, and coordinators' chances are 20 times better if they are coordinators in parallel projects in the framework program. Interestingly, the coefficient, which indicates how many programs the given organization participated in on average, was negative (PROG). This result suggests that winning new projects is not necessarily based on the results of previous similar projects. In other words, winning projects are not necessarily an advantage when they are a continuation of a previous project. This result apparently contradicts the results of LDA. Nevertheless, at the same time, the LDA showed that it is more common for participating partners to collaborate in projects that build on the results of a previous project. The coefficient of the former collaboration in the previous program had the greatest value; however, because of the low occurrence rate, this

Table 6
Group mean in LDA, the exponent of coefficients (Exp(β)), and significances in LogR. (Variable groups: c: Coorporate; C: Collaboration; E: Economy; T: Technology).

	Partners	Group mean (LDA)		LogR	Coordinators	Group mean (LDA)		LogR
		0	1	Exp(β)		0	1	Exp(β)
Coorporate (c)	FROM_TA	−0.0943	0.1243	1.0000**	TO_TA	−0.0980	0.1291	1.0000 **
	FROM_SR	0.0357	−0.0470	0.9986***	TO_SR	0.0354	−0.0467	0.9986 ***
	FROM_SH	−0.0946	0.1246	1.0000**	TO_SH	−0.0997	0.1313	1.0000 **
	FROM_RB	−0.0309	0.0407	1.0005***	TO_RB	−0.0313	0.0412	1.0005 ***
	FROM_RCB	−0.0284	0.0375	1.0004	TO_RCB	−0.0297	0.0392	1.0004
	FROM_PM	−0.0382	0.0503	1.0020***	TO_PM	−0.0407	0.0537	1.0021 ***
	FROM_PLF	−0.0851	0.1122	1.0000	TO_PLF	−0.0881	0.1160	1.0000
	FROM_PLB	−0.0852	0.1123	1.0000	TO_PLB	−0.0875	0.1153	1.0000
	FROM_OR	−0.1036	0.1365	1.0000*	TO_OR	−0.1075	0.1417	1.0000
	FROM_FA	−0.0870	0.1146	1.0000***	TO_FA	−0.0908	0.1197	1.0000 **
	FROM_EN	−0.1079	0.1421	1.0000	TO_EN	−0.1144	0.1507	1.0000
	FROM_CR	0.0177	−0.0233	0.9996	TO_CR	0.0170	−0.0224	0.9981
	FROM_CF	−0.0837	0.1103	1.0000	TO_CF	−0.0880	0.1160	1.0000
	FROM_Size	−0.1702	0.2242	1.1938***	TO_Size	−0.1831	0.2413	1.2054 ***
(E)	FROM_GDP	−0.0863	0.1137	1.0000	TO_GDP	−0.0937	0.1234	1.0000
	FROM_CI	−0.0352	0.0464	0.9941***	TO_CI	−0.0368	0.0485	0.9938 ***
	FROM_Pov	0.0195	−0.0257	1.0008	TO_Pov	0.0185	−0.0244	1.0013
(T)	FROM_PI	−0.1063	0.1400	1.0000	TO_PI	−0.1127	0.1485	1.0000
(C)	FROM_EC	−0.1725	0.2273	1.0000***	TO_EC	−0.1831	0.2412	1.0000 ***
	FROM_MULTI	−0.1857	0.2446	17.0223***	TO_MULTI	−0.1945	0.2562	20.1564 ***
	FROM_PROG	−0.0845	0.1113	0.5638***	TO_PROG	−0.0880	0.1160	0.5737 ***
(C)	Edges	0	1	Exp(β)				
	Dist	0.1142	−0.1504	0.9998***				
	FP7	−0.0413	0.0544	72.2008				

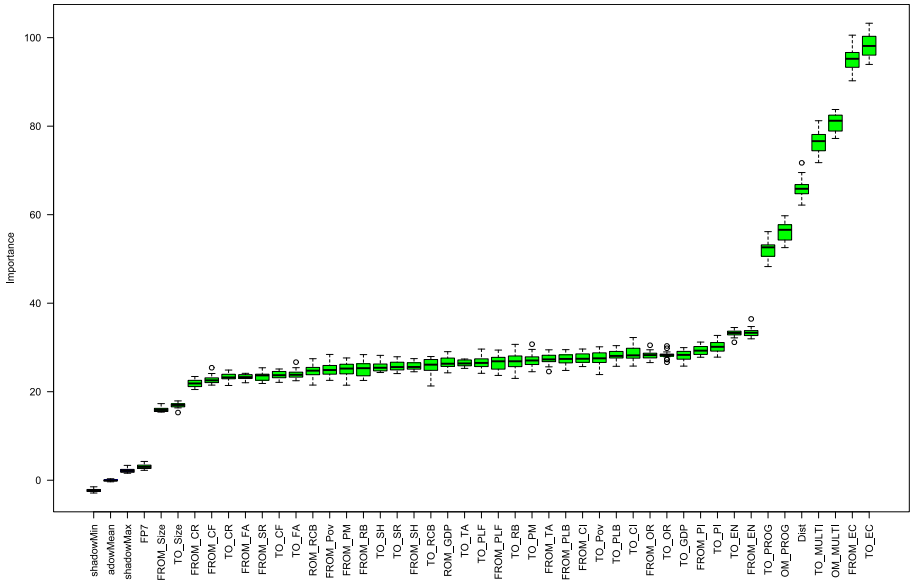


Fig. 2. Feature importance calculated by Boruta.

result was insignificant. The best generic method was the LogR method, with an accuracy of 0.83 and an F1 score of 0.90. However, similar to LDA and QDA, this method cannot handle the multicollinearity of explanatory variables. Therefore, nongeneric, black-box machine learning methods have also been employed.

3.3. Importance analysis of the outcome of random forest-based methods

To provide importance values for link prediction, RF methods are used. For variable selection, the Boruta method was used. After feature selection the RF and after the fine-tuning of parameters, a tuned RF method is employed.

3.3.1. Importance of features

Fig. 2 shows the results of Boruta feature selection.

In contrast to the LogR results, Boruta identified all the indicators as necessary because they are significantly greater than random (so-called shadow) variables. Boruta identified the most crucial variables, namely, the earned contribution (EC) and the average rate of parallel completion of projects (MULTI). This result reflects known preferential attachment or cumulative advantage. In other words, organizations that are already well connected, have earned, and run several projects simultaneously tend to receive new connections at a higher rate than organizations with fewer connections. It is challenging for newcomers to collaborate, and this variable is more challenging to overcome than geographical distance, which is the fifth most important indicator. The sixth and seventh most important indicators were still collaboration indicators (PROGs). This finding shows that for collaboration, former experience with successfully completed projects may be important. The results of LogR showed that previous experience does not necessarily

Table 7
Results of link prediction on the test dataset (T: Tuned method).

	LDA	QDA	LogR	SVM	SVM (T)	RF	RF (T)	XGBoost	XGBoost (T)
Accuracy	0.771900	0.817738	0.832748	0.844329	0.820240	0.856434	0.859057	0.853206	**0.863374**
Kappa	0.379527	0.371745	0.472292	0.535281	0.498282	0.587077	0.610799	0.580088	**0.620298**
Accuracy (Lower)	0.766624	0.812874	0.828043	0.839755	0.815401	0.852006	0.854662	0.848738	**0.859035**
Accuracy (Upper)	0.777112	0.822527	0.837376	0.848822	0.825003	0.860778	0.863367	0.857590	**0.867627**
Accuracy (Null)	0.767704	0.767704	0.767704	0.767704	0.767704	0.767704	0.767704	0.767704	0.767704
Accuracy PValue	0.059524	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
McNemar PValue	0.000000	0.000000	0.000000	0.000000	0.267566	0.000000	0.000020	0.000000	0.012707
Sensitivity	0.838116	**0.962262**	0.939399	0.924997	0.880952	0.917429	0.901556	0.913014	0.907180
Specificity	0.553066	0.340108	0.480285	0.577731	0.619594	0.654855	**0.718603**	0.655550	**0.718603**
Pos pred value	0.861062	0.828154	0.856602	0.878632	0.884439	0.897799	0.913706	0.897541	**0.914195**
Neg pred value	0.508301	**0.731689**	0.705717	0.699769	0.611626	0.705860	0.688353	0.695156	0.700830
Precision	0.861062	0.828154	0.856602	0.878632	0.884439	0.897799	0.913706	0.897541	**0.914195**
Recall	0.838116	**0.962262**	0.939399	0.924997	0.880952	0.917429	0.901556	0.913014	0.907180
F1	0.849434	0.890185	0.896092	0.901219	0.882692	0.907508	0.907590	0.905211	**0.910674**
Prevalence	0.767704	0.767704	0.767704	0.767704	0.767704	0.767704	0.767704	0.767704	0.767704
Detection rate	0.643425	**0.738732**	0.721180	0.710124	0.676310	0.704313	0.692128	0.700924	0.696445
Detection prevalence	0.747246	0.892023	0.841908	0.808215	0.764677	0.784489	0.757495	0.780939	0.761813
Balanced accuracy	0.695591	0.651185	0.709842	0.751364	0.750273	0.786142	0.810080	0.784282	**0.812892**

help the continuation of the given research, but those of LDA showed that the participation rate in such programs is greater than that in collaborating organizations. Only from the eighth indicator can we find a variable that characterizes the corporate, technological, or economic background. The most important variable among the corporate variables was the number of employees (EN), which refers to resource demand. The number of patents (PI) characterizes the technological background as a proxy variable, and the GDP per capita indicates the economic background. The last relevant indicator was previous collaborations between organizations in the former framework program (FP7). However, Boruta found that no features were unimportant.

Fig. 3 shows the importance values of indicators based on RF without parameter tuning (Fig. 3(a)) and with parameter tuning (Fig. 3(b)). Different types of indicators are marked with different colors. The importance of the variables is calculated by the mean decrease in the Gini index, which represents the average reduction in the Gini impurity index achieved by a particular feature when it is selected to split a node in a tree.

In line with the results of Boruta (see Fig. 2), Fig. 3 shows that the first seven most important values in link prediction are collaboration indicators. In addition to EC, the completion structure indicators (MULTI and PROG) play the most important role in forming new collaborations. These values are more important than geographical distance. Although the tuning of the RF method barely improved the accuracy and F1 score values, the importance of the variables somewhat changed, and the results were closer to the results of Boruta (compare Figs. 2 and 3(b)). After parameter tuning RF, the most important corporate indicators were the number of employees (EN), the GDP/capita, and the number of patents (PI). Interestingly, the number of employees (EN) played the greatest role in terms of the corporate indicators, and the size of the organization was less important within the corporate indicators. The results highlight that R&D&I projects are usually human resource intensive; therefore, organizations with sufficient human resources have a greater chance of participating in a collaboration.

3.4. Results of the black-box prediction methods

Even though the RF method can be considered a black-box method, the importance of the variables can still be interpreted. On the other hand, if the goal is only to maximize the accuracy of edge estimation and if it is important to use robust methods that can be successfully applied even if fewer conditions are met, then the use of nongeneric, black-box methods can also be justified. Table 7 shows the accuracy of the generic (LDA, QDA, and LogR) methods and the nongeneric methods, such as the RF, SVM, and XGBoost methods, on the test dataset.

Table 7 shows that the highest accuracy achieved by a method is 0.863374 with the tuned version of XGBoost, followed closely by the tuned version of RF, with an accuracy of 0.859057. These two methods also had the highest kappa values, indicating good agreement between the predicted and actual values. In terms of sensitivity, or the ability to correctly identify true positives, QDA and SVM (T) performed best, with values of 0.962262 and 0.924997, respectively. However, these methods yield lower specificity or the ability to correctly identify true negatives. Tuned RF and tuned XGBoost achieved the highest positive prediction, or the proportion of true positives among all positive predictions, and high precision, recall, and F1 scores. In summary, tuned RF and tuned XGBoost appear to be the best methods for link prediction in this collaboration network, as they achieved the highest accuracy and balanced accuracy while also achieving high precision, recall, and F1 scores. Nevertheless, all methods yielded a significantly higher accuracy than the null accuracy, which is the accuracy of a classifier that always predicts the most frequent class, 0.767704. Therefore, all methods have some predictive power for the link prediction task. The relatively low performance of LDA from generic algorithms and SVM from nongeneric algorithms suggests that there are nonlinear connections between the explanatory and explained variables. SVM is tested via several kernel functions. The best accuracy was achieved by the radial kernel and C-classification methods. The tuning process was only able to improve each link prediction method to a minimal extent.

3.5. Comparison of the prediction performance for network structure

The performance and accuracy of the link prediction models are also tested for the entire network. Table 8 shows the prediction errors of the structural parameters.

Table 8 shows that QDA yielded the lowest accuracy for the entire network. The accuracy values of all the generic methods were lower than 0.78. All methods overestimated the number of links because the difference between the original (G) and predicted ($\hat{G}$) links was negative. Therefore, each link prediction method predicted fewer components and denser networks. The structural parameters of the predicted networks were very close to those of the original networks. However, each prediction method predicted a less disassortative and less concentrated network, which is indicated by the negative differences in assortativity and positive differences in centrality. The average absolute value of differences between DC values indicates that the best (RF and XGBoost) link prediction methods can estimate the number of partners within ± one partner for organizations. The largest difference was observed in the estimation of the normalized BC. Here, the largest value was achieved by LDA, with a value of 0.202733. Within a collaboration community, the eigenvector centralities were well predicted by each link prediction method. The greatest average absolute difference

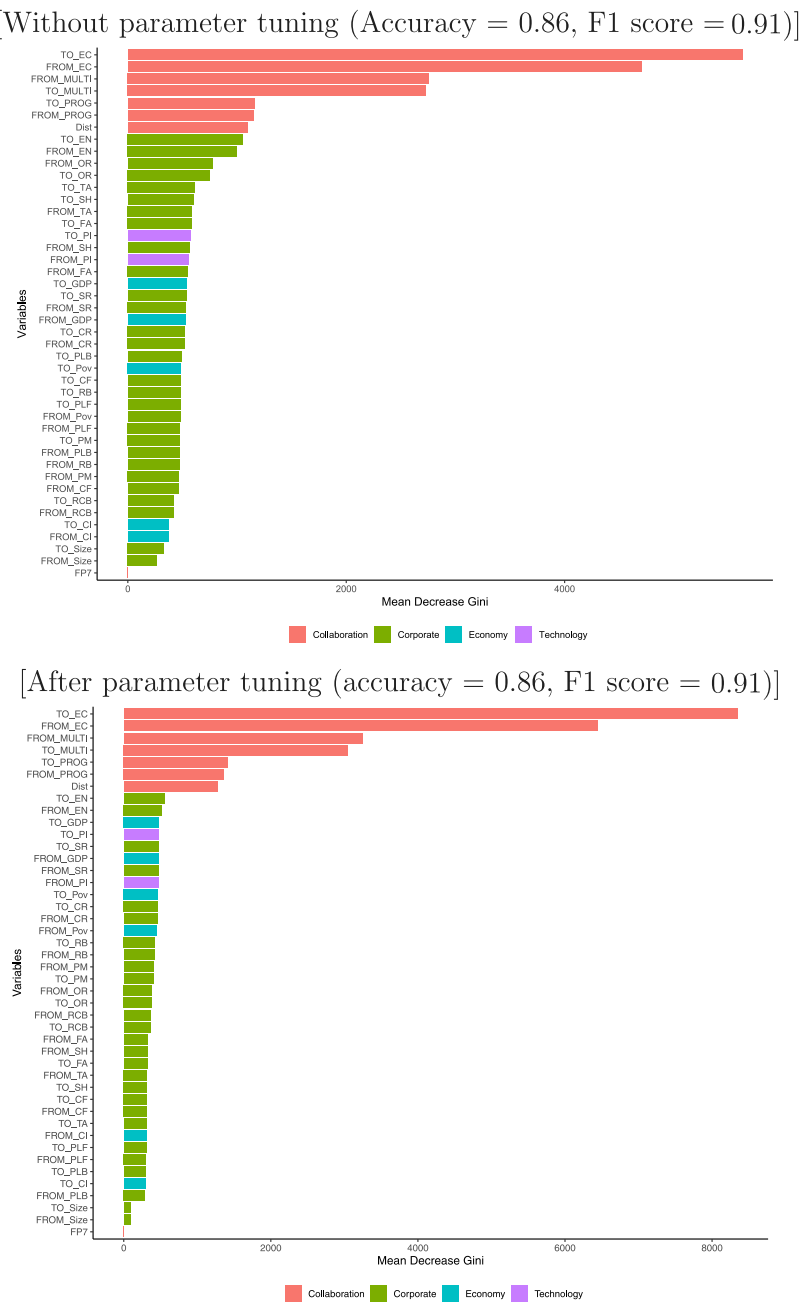


Fig. 3. Importance values of the RF methods.

Table 8
Accuracy of link prediction models for entire collaboration networks and the differences of network properties between the original (G) and predicted networks ($\hat{G}$).

	Network	LDA	QDA	LogR	SVM	SVM (T)	RF	RF (T)	XGBoost	XGBoost (T)
	Accuracy (all)	0.718438	0.710722	0.775996	0.834299	0.875079	0.965371	0.965989	0.965583	0.928062
Differences of network properties	Nodes	0	0	0	0	0	0	0	0	0
	Edges	−15208	−25859	−15199	−7349	−2514	−634	**−37**	−257	65
	Number of components	72	80	61	59	39	6	5	**4**	12
	Assortativity	−0.011348	−0.012460	−0.010368	−0.007606	−0.006570	−0.001907	**−0.001688**	−0.001850	−0.003846
	Density	−0.000037	−0.000064	−0.000037	−0.000018	−0.000006	−0.000002	**0.000000**	−0.000001	**0.000000**
	Average path length	−0.133734	−0.076950	−0.090980	−0.118293	−0.111480	**−0.023049**	−0.026438	−0.027690	−0.068630
	Degree centralization	0.016423	0.005245	0.010920	0.014792	0.015103	**0.003745**	0.004784	0.004835	0.011725
	Betweenness centralization	0.008601	−0.000696	0.002107	0.008538	0.008352	**0.001353**	0.002279	0.002573	0.007204
	Eigenvector centralization	0.001202	0.001095	0.000603	0.000842	0.000850	**0.000205**	0.000234	0.000231	0.000568
	Modularity	−0.009708	−0.012783	−0.012964	−0.002644	−0.006250	−0.002112	0.002740	−0.001129	**−0.000958**
Avg abs. diff.	Degree centrality	2.338588	2.835713	2.069601	1.485227	1.074658	0.320742	**0.304680**	0.325104	0.648721
	Normalized betweenness centrality	0.202733	0.145354	0.140855	0.170489	0.158315	**0.038794**	0.043452	0.046068	0.110242
	Eigenvector centrality	0.001912	0.001736	0.001493	0.001430	0.001455	**0.000483**	0.000510	0.000520	0.001088

Table 9
Top collaborators by their degree of centrality (links) and predictions based on generic (LDA, QDA, LogR) and non-generic methods (T: Tuned method).

Ord.	Org. ID	Org. Name	Orig. DC	LDA	QDA	LogR	SVM	SVM (T)	RF	RF (T)	XGBoost	XGBoost (T)
1	DE8170003303	Fraunhofer-Gesellschaft zur Foerderung der Angewandten Forschung eingetragener Verein	4918	4257	4709	4479	4322	4309	**4767**	4725	4723	4445
2	BE0419052173	Interuniversitair Micro-electronica Centrum	1870	1471	1736	1567	1499	1474	**1761**	1744	1739	1565
3	DK30060946	Danmarks Tekniske Universitet	1687	1464	1602	1552	1497	1470	**1627**	1620	1620	1532
4	BE0425260668	Katholieke Universiteit te Leuven	1455	1449	1454	1451	1451	1450	**1455**	1454	1454	1453
5	DE5030021537	Deutsches Zentrum Fuer Luftund Raumfahrt E.V.	1372	1198	1316	1258	1216	1203	**1335**	1326	1318	1261
6	DK31119103	Aarhus Universitet	1233	1137	1172	1130	1127	1114	1186	1180	**1187**	1157
7	NL09098104	Stichting Wageningen Research	1200	1103	1167	1115	1107	1107	**1185**	1179	1182	1134
8	NO919303808	SINTEF AS	1191	1094	1158	1108	1097	1087	**1167**	1162	1162	1124
9	BE0248015142	Universiteit GENT	1150	1064	1146	1104	1087	1083	**1123**	1115	1114	1110
10	AT9110045499	AIT Austrian Institute of Technology GMBH	1115	1054	1111	1046	1050	1047	**1115**	1114	1112	1073

was lower than 0.2%. Table 8 indicates that tuning the parameters in nongeneric methods no longer significantly improves link prediction. Moreover, the estimation of certain structural parameters, such as centralizations, may even deteriorate.

Since the link prediction accuracies not only on the test and training sets but also on the entire network were acceptable and since the differences in the structural parameters were low, it is possible to predict collaborations and which organizations play a key role in the collaboration network. The accuracy of the best nongeneric method for all links was 0.928, which means that 92.8% of the link predictions were correct. Table 9 shows the results for the top collaborative organizations by predicting the degree of centrality.

Table 9 shows that the top collaborators are higher education institutes, such as Denmark University of Technology, Aarhus University, University of Gent, and University of Lauven; research institutes, such as Stichting Wageningen Research and Interuniversitair Micto-electronica Centrum; and innovative technology-oriented companies, such as Fraunhofer-Gesellschaft zur Foerderung der Angewandten Forschung e. V. (Fraunhofer) and the AIT Austrian Institute of Technology GMBH. While each method underestimated the number of links, the best estimation was provided by the RF methods. The largest DC value was for the German company Fraunhofer-Gesellschaft zur Foerderung der Angewandten Forschung e. V. (Fraunhofer), which has two collaboration partners as the second organization. Fraunhofer has approximately 30,000 employees. This is the largest organization for applied research and development services in Europe. It represents an important part of the German research landscape and consists of colleges (especially universities), the Max Planck Society, the Helmholtz Association of German Research Centers, the Gottfried Wilhelm Leibniz Scientific Association and the German Research Association. The seat is in Munich, Germany. The remaining top collaborators are also related to the universities. A university is directly or indirectly behind most nonhigher education institutions as the dominant owner.

3.6. Detecting collaboration communities

Collaboration communities provide a set of organizations where collaboration is stronger than the estimated link prediction model. The best link prediction method (RF) underestimated links by only 1618 cases for 1198 collaborators. Fig. 4 shows the collaboration communities for these underestimated collaborations.

The collaboration communities contained seven components and 26 modules. Since the link prediction method already includes the distances between organizations as an explanatory variable, modules are already distance independent. However, Fig. 4 shows that most nodes are concentrated in core countries of the European Union, especially in Germany and the UK. The UK is tied to continental Europe by hundreds of links, which Brexit will significantly complicate in the forthcoming period. Different node colors represent different modules. The largest (red) modules are mostly in Germany and the UK. Moreover, the modules cannot be linked to a single administrative area. Communities in newcomer countries are more heterogeneous, and there are fewer underestimated collaborators and collaborations.

4. Discussion

The different methods all confirm that the collaboration network is quite extensive. The H2020 collaboration network is highly concentrated but fragmented, with very few intermediate connecting communities (see Table 4). Moreover, this collaboration network structure can support link prediction methods (Zhou, Lü, & Zhang, 2009). Almost half of the organizations that have an identifier (BvdId) in ORBIS and thus also have access to corporate data are large or very large companies (see Table 5). These companies are much more capital intensive than small and medium-sized enterprises are, the support of which is an important objective of H2020. Although Fig. 2 highlights that all the incorporated variables are important in the link prediction model, the results clearly show that organizations with strong capital and a good economic and technological background are more likely to cooperate (see Table 6 and compare them with Fig. 3). These results are even

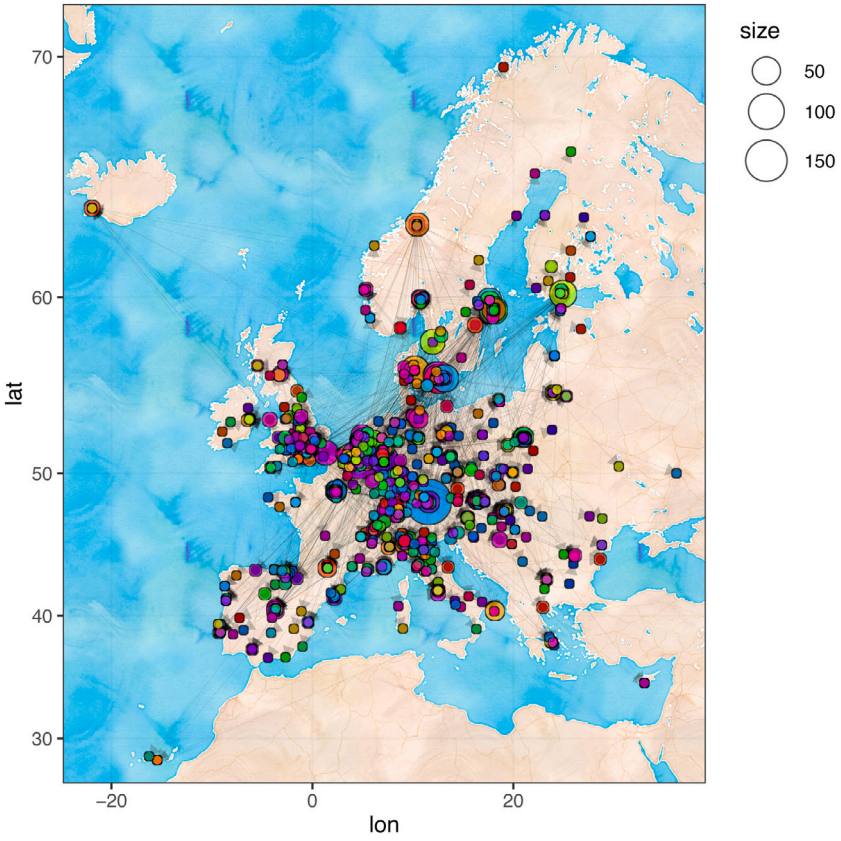


Fig. 4. Economic communities. Nodes represent organizations, and links between nodes (from partners to coordinator(s)) represent underestimated collaborations. The size of each node follows the logarithm of the logarithm of the DC values. Different colors represent different modules.

more important for coordinators. Participating in several simultaneously running projects (see the role of the indicator MULTI) or having previously successful projects (see the role of the indicator PROG) increases the chance of collaboration.

This result explains the high concentration, fragmentation, and formation of small collaborative communities. We expected similar results based on previous research (see e.g., Roediger-Schluga & Barber, 2008; Scherngell & Lata, 2013); however, these two indicators (MULTI and PROG), in addition to the earned contribution (EC), essentially dominate the development of collaborative relationships (see Fig. 3). These novel findings were achievable due to the use of a comprehensive dataset that includes multiple sources, such as ORBIS, CORDIS, PATSTAT and EUROSTAT, enabling the analysis of new variables, the detection of new relationships and the achievement of high prediction accuracy.

In addition, both generic methods, such as the LDA, QDA, and LogR methods, and nongeneric black-box methods, such as the SVM, RF, and XGBoost methods, can predict links between collaborators. On the one hand, nongeneric methods achieved better precision (see Table 7); on the other hand, only generic methods and the RF method supported this interpretation. Tuned nongeneric methods only slightly improved link prediction (see Tables 7 and 8). Testing the aforementioned methods resulted in remarkable improvements in the prediction performance achieved by previous studies available in the literature on collaboration analysis (Chen et al., 2021; Wang et al., 2015).

The analysis of measures employed in current collaborations indicates that these collaborations exhibit a certain degree of predictability, suggesting that they are relatively fixed and stable. This can have both advantageous and detrimental consequences for collaboration. Foreseeable collaborations can serve as an indication of enduring and firmly formed connections between businesses, which can be advantageous for enduring partnerships and consistency in research and innovation endeavors. Having the ability to determine which collaborations are likely to be successful can assist organizations in effectively distributing resources, prioritizing the enhancement of current partnerships rather than continually pursuing new partnerships. Foreseeable partnerships may indicate a degree of confidence and reciprocal comprehension among collaborators, resulting in more seamless communication, shared objectives, and triumphant results. Nevertheless, established partnerships can restrict the potential for investigating novel concepts, technologies, or methodologies. Organizations that fail to actively seek diverse alliances may become complacent and thus miss possible breakthroughs. When collaborations become overly predictable, businesses may become less willing to take chances or explore new domains, which could impede innovation and impair growth. Established partnerships may result in a dearth of diversity in viewpoints, specialized knowledge, and assets, all of which are crucial for addressing intricate problems and fostering ingenuity.

Finally, the results showed that collaboration is self-concentrated. The dominance of strong organizations (those with good corporate, technological, or economic backgrounds) is high; however, compared to those of well-estimated link prediction models, in some cases, the actual number of connections is lower, and the real network is more fragmented and concentrated than the predicted subgraph (see Table 8). This result suggested that additional unexplored mechanisms escalated the increases in concentration and fragmentation. When each model underestimates links, the strongest, most connected organizations exist (see Table 9). These organizations specify a collaboration community geographically concentrated in the western half of Europe in Germany, Great Britain, Spain, and Italy (see Fig. 4). These novel and important findings are not observed without analyzing the outliers, unlike in general behavior, indicating the importance of analyzing the relevant weak signals in collaboration network analysis.

Table 10
Summary of conclusions, contributions and implications.

Nr	Result	Conclusion	Contribution	Implication
1	A highly accurate link prediction model is proposed.	Two generic and three nongeneric methods proved to be applicable for the link prediction problem. The nongeneric methods achieved higher accuracy, suggesting the existence of nonlinear relationships across the model. Among the nongeneric methods, the random forest-based methods proved to be the most effective.	The evaluation of the models helps scholars achieving efficient model selection in collaboration analysis.	The best performing model can be used in the next FPs to predict the formation of consortia.
2	The determining factors of collaborations were identified by using diverse databases.	The most important factors were those related to the embeddedness in former collaborations.	We showed that to build a highly accurate collaboration link prediction model, the consideration of the embeddedness-related variables is extremely important.	A ranking of the influencing factors was provided. This ranking is useful for the decision-makers in supporting the collaboration formation in future FPs.
3	Outliers, i.e., relationships with stronger than predicted connections, were detected.	These connections were concentrated in organizations from core countries.	We proposed economic null model to specify collaboration communities, where the collaboration stronger than the prediction of link prediction model.	In the Framework Programmes, this information provides opportunities for stakeholders to analyze concentrations and supports the control (balance or further strengthening of concentrated collaborations) if necessary.

5. Summary and conclusion

In this paper, we proposed an accurate link prediction model to analyze and predict H2020 collaboration by combining multiple databases and extending the analyzed variable set with new factors. This paper has several contributions and implications, which are presented in Table 10.

In this paper, a highly accurate link prediction model is proposed. Two generic and three nongeneric methods were investigated, of which the nongeneric methods were proven to be more accurate, indicating that nonlinear relationships play a significant role in the models. For decision-makers, we demonstrated the model accuracy that can be achieved, which is beneficial for predicting and developing future consortia in FPs. This study supported effective model selection during the analysis.

In addition, we identified and ranked the key factors driving the collaboration of organizations. Embeddedness in previous projects, such as multiproject and program characteristics and earned contributions, essentially determines the likelihood of collaboration. These results can help decision-makers effectively support the formation of future FP collaborations.

Finally, the study specified the outlier collaboration communities, which are limited to the organizations of the EU core countries. The identification of these weak yet significant observations raises attention to a deeper analysis of core countries' collaboration as a new research direction. This finding provides opportunities for stakeholders to analyze collaboration concentrations in FPs and either strengthen existing collaborations or balance them by involving organizations outside the core countries.

Funding

This research is supported by the Research Centre at the Faculty of Business and Economics (PE-GTK-GSKK A095000000-7) of the University of Pannonia (Veszprém, Hungary). Project no. K 142395 has been implemented with support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the K_22 "OTKA" funding scheme.

Project no. KDP-11-3/PALY-2021 has been implemented with support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the KDP-2020 funding scheme.

CRediT authorship contribution statement

Zsolt T. Kosztyán: Conceptualization, Methodology, Software, Writing – review & editing. **Ferenc Király:** Data curation, Writing – original draft. **Attila I. Katona:** Visualization, Investigation, Writing – review & editing. **Tibor Csizmadia:** Investigation, Writing – review & editing. **Beáta Fehérvölgyi:** Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2024.124417.

References

Adafre, S. F., & de Rijke, M. (2005). Discovering missing links in wikipedia. In *Proceedings of the 3rd international workshop on link discovery* (pp. 90–97). http://dx.doi.org/10.1145/1134271.1134284.

Ahmad, I., Basheri, M., Iqbal, M. J., & Rahim, A. (2018). Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. *IEEE Access*, *6*, 33789–33795. http://dx.doi.org/10.1109/ACCESS.2018.2841987.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., Xing, E. P., & Jaakkola, T. (2006). Mixed membership stochastic block models for relational data with application to protein-protein interactions. In *Proceedings of the international biometrics society annual meeting* (p. 1). http://dx.doi.org/10.1007/978-3-540-73133-7_5.

Alibrahim, H., & Ludwig, S. A. (2021). Hyperparameter optimization: Comparing genetic algorithm against grid search and bayesian optimization. In *2021 IEEE congress on evolutionary computation* (pp. 1551–1559). IEEE, http://dx.doi.org/10.1109/CEC45853.2021.9504761.

Bastami, E., Mahabadi, A., & Taghizadeh, E. (2019). A gravitation-based link prediction approach in social networks. *Swarm and Evolutionary Computation*, *44*, 176–186. http://dx.doi.org/10.1016/j.swevo.2018.03.001.

Cai, Y. (2023). Towards a new model of eu-china innovation cooperation: Bridging missing links between international university collaboration and international industry collaboration. *Technovation*, *119*, Article 102553. http://dx.doi.org/10.1016/j.technovation.2022.102553.

Camarinha-Matos, L. M., Afsarmanesh, H., Galeano, N., & Molina, A. (2009). Collaborative networked organizations–concepts and practice in manufacturing enterprises. *Computers & Industrial Engineering*, *57*, 46–60. http://dx.doi.org/10.1016/j.cie.2008.11.024.

Camarinha-Matos, L. M., Fornasiero, R., & Afsarmanesh, H. (2017). Collaborative networks as a core enabler of industry 4.0. In *Collaboration in a data-rich world: 18th IFIP WG 5.5 working conference on virtual enterprises, PRO-VE 2017, vicenza, Italy, September (2017) 18-20, proceedings 18* (pp. 3–17). Springer, http://dx.doi.org/10.1007/978-3-319-65151-4_1.

Chang, H. w., & Huang, M. H. (2013). Prominent institutions in international collaboration network in astronomy and astrophysics. *Scientometrics*, *97*, 443–460. http://dx.doi.org/10.1007/s11192-013-0976-x.

Chen, W., Qu, H., & Chi, K. (2021). Partner selection in china interorganizational patent cooperation network based on link prediction approaches. *Sustainability*, *13*(1003), http://dx.doi.org/10.3390/su13021003.

Chesbrough, H. W. (2003). *Open innovation: the new imperative for creating and profiting from technology*. Harvard Business Press, http://dx.doi.org/10.1016/j.jengtecman.2004.05.003.

Chi, K., Yin, G., Dong, Y., & Dong, H. (2019). Link prediction in dynamic networks based on the attraction force between nodes. *Knowledge-Based Systems*, *181*, Article 104792. http://dx.doi.org/10.1016/j.knosys.2019.05.035.

Chuanming, Y., Yutian, G., Xiaoli, Z., & Lu, A. (2017). Collaboration recommendation of finance research based on multi-feature fusion. *Data Analysis and Knowledge Discovery*, *1*, 39–47. http://dx.doi.org/10.11925/infotech.2096-3467.2017.08.05.

Cséfalvay, Z., & Gkotsis, P. (2022). Robotisation race in europe: the robotisation chain approach. *Economics of Innovation and New Technology*, *31*, 693–710. http://dx.doi.org/10.1080/10438599.2020.1849968.

De Noni, I., Orsi, L., & Belussi, F. (2018). The role of collaborative networks in supporting the innovation performances of lagging-behind european regions. *Research Policy*, *47*, 1–13. http://dx.doi.org/10.1016/j.respol.2017.09.006.

De Prato, G., & Nepelski, D. (2014). Global technological collaboration network: Network analysis of international co-inventions. *The Journal of Technology Transfer*, *39*, 358–375. http://dx.doi.org/10.1007/s10961-012-9285-4.

Demir, S., & Şahin, E. K. (2022). Liquefaction prediction with robust machine learning algorithms (svm, rf, and xgboost) supported by genetic algorithm-based feature selection and parameter optimization from the perspective of data processing. *Environmental Earth Sciences*, *81*, 459. http://dx.doi.org/10.1007/s12665-022-10578-4.

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2006). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *19*, 1–16. http://dx.doi.org/10.1109/TKDE.2007.250581.

Enkel, E., Gassmann, O., & Chesbrough, H. (2009). Open R & D and open innovation: exploring the phenomenon. *R & D Management*, *39*, 311–316. http://dx.doi.org/10.1111/j.1467-9310.2009.00570.x.

Gadar, L., Kosztyan, Z. T., & Abonyi, J. (2018). The settlement structure is reflected in personal investments: distance-dependent network modularity-based measurement of regional attractiveness. *Complexity*, *2018*, 1–16. http://dx.doi.org/10.1155/2018/1306704.

Ge, X., Wang, Y. C., Wang, B., Kuo, C. C. J., et al. (2024). Knowledge graph embedding: An overview. *APSIPA Transactions on Signal and Information Processing*, *13*, http://dx.doi.org/10.1561/116.00000065.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*, 7821–7826. http://dx.doi.org/10.1073/pnas.122653799.

Gold, C., & Sollich, P. (2003). Model selection for support vector machine classification. *Neurocomputing*, *55*, 221–249. http://dx.doi.org/10.1016/S0925-2312(03)00375-8.

Gu, Q., Li, Z., & Han, J. (2011). Linear discriminant dimensionality reduction. In *Machine learning and knowledge discovery in databases: European conference, ECML pKDD 2011, athens, Greece, September (2011) 5-9. proceedings, part i 11* (pp. 549–564). Springer, http://dx.doi.org/10.1007/978-3-642-23780-5_45.

Guan, Q., An, H., Gao, X., Huang, S., & Li, H. (2016). Estimating potential trade links in the international crude oil trade: A link prediction approach. *Energy*, *102*, 406–415. http://dx.doi.org/10.1016/j.energy.2016.02.099.

Guan, J., Zuo, K., Chen, K., & Yam, R. C. (2016). Does country-level R & D efficiency benefit from the collaboration network structure?. *Research Policy*, *45*, 770–784. http://dx.doi.org/10.1016/j.respol.2016.01.003.

Guns, R., & Rousseau, R. (2014). Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics*, *101*, 1461–1473. http://dx.doi.org/10.1007/s11192-013-1228-9.

Guns, R., & Wang, L. (2017). Detecting the emergence of new scientific collaboration links in africa: A comparison of expected and realized collaboration intensities. *Journal of Informetrics*, *11*, 892–903. http://dx.doi.org/10.1016/j.joi.2017.07.004.

Han, P., Shi, J., Li, X., Wang, D., Shen, S., & Su, X. (2014). International collaboration in lis: global trends and networks at the country and institution level. *Scientometrics*, *98*, 53–72. http://dx.doi.org/10.1007/s11192-013-1146-x.

Hasan, M. A., & Zaki, M. J. (2011). A survey of link prediction in social networks. *Social Network Data Analytics*, *24*, 3–275. http://dx.doi.org/10.1007/978-1-4419-8462-3_9.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, *13*, 18–28. http://dx.doi.org/10.1109/5254.708428.

Huang, L., Zhu, Y., Zhang, Y., Zhou, X., & Jia, X. (2018). A link prediction-based method for identifying potential cooperation partners: A case study on four journals of informetrics. In *2018 Portland international conference on management of engineering and technology* PICMET, (pp. 1–6). IEEE, http://dx.doi.org/10.23919/PICMET.2018.8481974.

Kosztyán, Z. T., Csányi, V. V., Banász, Z., Jakobi, Á., Neumanné-Virág, I., & Telcs, A. (2021). The role of higher education in spatial mobility. *Applied Network Science*, *6*, 1–30. http://dx.doi.org/10.1007/s41109-021-00428-w.

Kosztyán, Z. T., Katona, A. I., Kuppens, K., Kisgyörgy-Pál, M., Nachbagauer, A., & Csizmadia, T. (2022). Exploring the structures and design effects of eu-funded R & D & i project portfolios. *Technological Forecasting and Social Change*, *180*, Article 121687. http://dx.doi.org/10.1016/j.techfore.2022.121687.

Kosztyán, Z. T., Király, F., & Kurbucz, M. T. (2022). Analysis of ownership network of european companies using gravity models. *Applied Network Science*, *7*, 1–31. http://dx.doi.org/10.1007/s41109-022-00501-y.

Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta–A system for feature selection. *Fundamenta Informaticae*, *101*, 271–285. http://dx.doi.org/10.3233/FI-2010-288.

Lande, D., Fu, M., Guo, W., Balagura, I., Gorbov, I., & Yang, H. (2020). Link prediction of scientific collaboration networks based on information retrieval. *World Wide Web*, *23*, 2239–2257. http://dx.doi.org/10.1007/s11280-019-00768-9.

Lee, D. H., Seo, I. W., Choe, H. C., & Kim, H. D. (2012). Collaboration network patterns and research performance: the case of korean public research institutions. *Scientometrics*, *91*, 925–942. http://dx.doi.org/10.1007/s11192-011-0602-8.

Li, T., Wang, B., Jiang, Y., Zhang, Y., & Yan, Y. (2018). Restricted boltzmann machine-based approaches for link prediction in dynamic networks. *IEEE Access*, *6*, 29940–29951. http://dx.doi.org/10.1109/ACCESS.2018.2840054.

Liu, S., Dong, Z., Ding, C., Wang, T., & Zhang, Y. (2020). Do you need cobalt ore? estimating potential trade relations through link prediction. *Resources Policy*, *66*, Article 101632. http://dx.doi.org/10.1016/j.resourpol.2020.101632.

Liu, Y., & Kou, Z. (2007). Predicting who rated what in large-scale datasets. *ACM SIGKDD Explorations Newsletter*, *9*, 62–65. http://dx.doi.org/10.1145/1345448.1345462.

Liu, Y., Zhao, C., Wang, X., Huang, Q., Zhang, X., & Yi, D. (2016). The degree-related clustering coefficient and its application to link prediction. *Physica A: Statistical Mechanics and its Applications*, *454*, 24–33. http://dx.doi.org/10.1016/j.physa.2016.02.014.

MacCarthy, B. L., & Jayarathne, P. (2012). Sustainable collaborative supply networks in the international clothing industry: a comparative analysis of two retailers. *Production Planning and Control*, *23*, 252–268. http://dx.doi.org/10.1080/09537287.2011.627655.

Mahnken, T. A., & Moehrle, M. G. (2018). Multi-cross-industry innovation patents in the usa-a combination of patstat and orbis search. *World Patent Information*, *55*, 52–60. http://dx.doi.org/10.1016/j.wpi.2018.10.003.

Malin, B., Airoldi, E., & Carley, K. M. (2005). A network analysis model for disambiguation of names in lists. *Computational & Mathematical Organization Theory*, *11*, 119–139. http://dx.doi.org/10.1007/s10588-005-3940-3.

Mohan, A., Venkatesan, R., & Pramod, K. (2017). A scalable method for link prediction in large real world networks. *Journal of Parallel and Distributed Computing*, *109*, 89–101. http://dx.doi.org/10.1016/j.jpdc.2017.05.009.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*, Article 026113. http://dx.doi.org/10.1103/PhysRevE.69.026113.

Novitzky, P., Bernstein, M. J., Blok, V., Braun, R., Chan, T. T., Lamers, W., et al. (2020). Improve alignment of research policy and societal values. *Science*, *369*, 39–41. http://dx.doi.org/10.1007/s11192-013-1146-x/10.1126/science.abb3415.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*, 217–222. http://dx.doi.org/10.1080/01431160412331269698.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). *Vol. 9*, *Hyperparameters and tuning strategies for random forest*. Wiley Interdisciplinary Reviews: data mining and knowledge discovery, Article e1301. http://dx.doi.org/10.1002/widm.1301.

Qi, Y., Zhang, X., Hu, Z., Xiang, B., Zhang, R., & Fang, S. (2022). Choosing the right collaboration partner for innovation: a framework based on topic analysis and link prediction. *Scientometrics*, *127*, 5519–5550. http://dx.doi.org/10.1007/s11192-022-04306-9.

Roediger-Schluga, T., & Barber, M. J. (2008). R & D collaboration networks in the european framework programmes: Data processing, network construction and selected results. *International Journal of Foresight and Innovation Policy*, *4*, 321–347. http://dx.doi.org/10.1504/IJFIP.2008.017583.

Rossi, A., Barbosa, D., Firmani, D., Matinata, A., & Merialdo, P. (2021). Knowledge graph embedding for link prediction: A comparative analysis. *Vol. 15*, In *ACM transactions on knowledge discovery from data* (pp. 1–49). http://dx.doi.org/10.1145/3424672.

Sachwald, F. (2008). Location choices within global innovation networks: the case of europe. *The Journal of Technology Transfer*, *33*, 364–378. http://dx.doi.org/10.1007/s10961-007-9057-8.

Sagi, O., & Rokach, L. (2018). *Vol. 8*, *Ensemble learning: A survey*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Article e1249. http://dx.doi.org/10.1002/widm.1249.

Scherngell, T., & Lata, R. (2013). Towards an integrated european research area? findings from eigenvector spatially filtered spatial interaction models using european framework programme data. *Papers in Regional Science*, *92*, 555–577. http://dx.doi.org/10.1111/j.1435-5957.2012.00419.x.

Secundo, G., Toma, A., Schiuma, G., & Passiante, G. (2019). Knowledge transfer in open innovation: A classification framework for healthcare ecosystems. *Business Process Management Journal*, *25*, 144–163. http://dx.doi.org/10.1108/BPMJ-06-2017-0173.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & De Freitas, N. (2015). Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, *104*, 148–175. http://dx.doi.org/10.1109/JPROC.2015.2494218.

Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J., & Gifford, E. M. (2016). Extreme gradient boosting as a method for quantitative structure–activity relationships. *Journal of Chemical Information and Modeling*, *56*, 2353–2360. http://dx.doi.org/10.1021/acs.jcim.6b00591.

Sun, Q., Hu, R., Yang, Z., Yao, Y., & Yang, F. (2017). An improved link prediction algorithm based on degrees and similarities of nodes. In *2017 IEEE/aCIS 16th international conference on computer and information science* ICIS, (pp. 13–18). IEEE, http://dx.doi.org/10.1109/ICIS.2017.7959962.

Tarasconi, G., & Menon, C. (2017). *Matching crunchbase with patent data*: *Technical Report*, OECD. fa-en, http://dx.doi.org/10.1787/15f967.

Tharwat, A. (2016). Linear vs. quadratic discriminant analysis classifier: a tutorial. *International Journal of Applied Pattern Recognition*, *3*, 145–180. http://dx.doi.org/10.1504/IJAPR.2016.079050.

Tonkin, M., Woodhams, J., Bull, R., Bond, J. W., & Santtila, P. (2012). A comparison of logistic regression and classification tree analysis for behavioural case linkage. *Journal of Investigative Psychology and Offender Profiling*, *9*, 235–258. http://dx.doi.org/10.1002/jip.1367.

Veugelers, R., Cincera, M., Frietsch, R., Rammer, C., Schubert, T., Pelle, A., et al. (2015). The impact of horizon 2020 on innovation in europe. *Intereconomics*, *50*, 4–30. http://dx.doi.org/10.1007/s10272-015-0521-7.

Wang, Z., Han, W., Sun, Z., & Pan, X. (2019). Research on scientific collaboration prediction based on the combination of network topology and node attributes. *Information Studies: Theory & Application*, *42*, 116–120.

Wang, P., Xu, B., Wu, Y., & Zhou, X. (2015). Link prediction in social networks: The state-of-the-art. *Science China. Information Sciences*, *58*, 1–38, URL http://dx.doi.org/10.1007/s11432-014-5237-y.

Xu, M., & Yin, Y. (2017). A similarity index algorithm for link prediction. In *2017 12th international conference on intelligent systems and knowledge engineering* ISKE, (pp. 1–6). IEEE, http://dx.doi.org/10.1109/ISKE.2017.8258724.

Yang, Q., Dong, E., & Xie, Z. (2014). Link prediction via nonnegative matrix factorization enhanced by blocks information. In *2014 10th international conference on natural computation* ICNC, (pp. 823–827). IEEE, http://dx.doi.org/10.1109/ICNC.2014.6975944.

Yu, D., Xu, Z., & Wang, X. (2020). Bibliometric analysis of support vector machines research trend: A case study in china. *International Journal of Machine Learning and Cybernetics*, *11*, 715–728. http://dx.doi.org/10.1007/s13042-019-01028-y.

Zhang, S., Tong, H., Xu, J., & Maciejewski, R. (2019). Graph convolutional networks: a comprehensive review. *Computational Social Networks*, *6*, 1–23. http://dx.doi.org/10.1186/s40649-019-0069-y.

Zhang, C., Zhang, H., Yuan, D., & Zhang, M. (2016). Deep learning based link prediction with social pattern and external attribute knowledge in bibliographic networks. *Vol. 81*, In *2016 IEEE international conference on internet of things (iThings) and IEEE green computing and communications (greenCom) and IEEE cyber, physical and social computing (cPSCom) and IEEE smart data* smartData, (pp. 5–821). IEEE, http://dx.doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2016.170.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., et al. (2020). Graph neural networks: A review of methods and applications. *AI Open*, *1*, 57–81. http://dx.doi.org/10.1016/j.aiopen.2021.01.001.

Zhou, T., Lü, L., & Zhang, Y. C. (2009). Predicting missing links via local information. *The European Physical Journal B*, *71*, 623–630. http://dx.doi.org/10.1140/epjb/e2009-00335-8.

Zhu, J., Hong, J., & Hughes, J. G. (2002). Using markov models for web site link prediction. In *Proceedings of the thirteenth ACM conference on hypertext and hypermedia* (pp. 169–170). http://dx.doi.org/10.1145/513338.513381.