Analysis of the evolution of scientific collaboration networks for the prediction of new co-authorships

Análise da evolução de redes de colaboração científica para a predição de novas coautorias

Felipe AFFONSO[1] 0000-0003-2986-2379

Monique de Oliveira SANTIAGO[1] 0000-0003-0163-2774

Thiago Magela RODRIGUES DIAS[1] 0000-0001-5057-9936

Abstract

When publishing an article with other authors, initial links must be formed by a collaboration between authors, a scientific collaboration network. In this context, the papers are represented by the edges, and the authors are represented the nodes, forming a network. At this moment, the following question arises: How does the evolution of the network occur over time? Understanding what factors are essential for creating a new connection to answer this question is necessary. Therefore, the purpose of this article is to foresee connections in co-authorship networks formed by PhDs with curricula registered in Lattes Platform in the areas of Information Sciences and Biology. The following steps are performed: initially the data is extracted and organized. This step is essential for the continuity of the process. Then, co-authorship networks are generated based on articles published together. Subsequently, the attributes to be used are defined and some metrics are calculated. Finally, machine learning algorithms estimate future scientific collaborations in the selected areas. The Lattes Platform has 6.6 million resumes for researchers and represents one of the most relevant and recognized scientific repositories worldwide. As a result, random forest and logistic regression algorithms showed the highest hit rates, and preferential attachment attribute was identified as the most influential in the emergence of new scientific collaborations. Through the results, it is possible to establish the evolution of the network of scientific associations of researchers at a national level, assisting development agencies in selecting of future outstanding researchers.

Keywords: Co-authorship networks. Lattes Platform. Scientific data repositories.

Resumo

Ao publicar um artigo em conjunto com outros autores, inicialmente deve-se formar vínculos pela colaboração entre eles, o que pode ser caracterizado como uma rede de colaboração científica. Nesse contexto, os trabalhos representam as arestas e os autores representam os nós, formando uma rede. Nesse momento surge a seguinte dúvida: Como a evolução da rede ocorre ao longo do tempo? Para responder a essa pergunta, é necessário entender quais fatores são essenciais para a criação de uma nova conexão. O objetivo deste artigo é prever conexões em redes de coautoria formadas por doutores com currículos registrados na Plataforma Lattes nas áreas de Ciências da Informação e Biologia. Para tanto, as seguintes etapas são executadas: inicialmente os dados são extraídos

[1] Centro Federal de Educação Tecnológica de Minas Gerais, Programa de Pós-Graduação em Modelagem Matemática e Computacional. Av. Amazonas, 7675, Nova Gameleira, 30510-000, Belo Horizonte, MG, Brasil. *Correspondência para*/Correspondence to: T. M. RODRIGUES DIAS. E-mail: <thiagomagela@gmail.com>.

Received on May 14, 2021, final version resubmitted on October 30, 2021 and approved on November 25, 2021.

Como citar este artigo/How to cite this article

Afonso, F.; Santiago, M. O.; Rodrigues Dias, T. M. Analysis of the evolution of scientific collaboration networks for the prediction of new co-authorships. *Transinformação*, v. 34, e200033, 2022. https://doi.org/10.1590/2318-0889202234e200033

CC BY

e organizados. Essa etapa é fundamental para a continuidade do processo. Em seguida, as redes de coautoria são geradas tomando como base artigos publicados em conjunto. Posteriormente, os atributos a serem utilizados são definidos e as métricas são calculadas. Por fim, algoritmos de aprendizado de máquinas são utilizados para estimar futuras colaborações científicas nas áreas selecionadas. Atualmente, a Plataforma Lattes possui 6,6 milhões de currículos de pesquisadores e representa um dos repositórios científicos mais relevantes e reconhecidos em todo o mundo. Como resultado, os algoritmos "florestas aleatórias" e "regressão logística" apresentaram as maiores taxas de acerto, e o atributo "atração preferencial" foi identificado como mais influente no surgimento de novas colaborações científicas. Através dos resultados, é possível estabelecer a evolução da rede de colaborações científicas de pesquisadores em nível nacional, auxiliando as agências de desenvolvimento na seleção de futuros pesquisadores destacados.

***Palavras-chave**: Redes de coautoria. Plataforma Lattes. Repositórios de dados científicos.*

Introduction

In the late 1990s, several researchers devoted attention to network studies. Work has been done on biology, the internet, routers, among others (Newman, 2001; Newman; Park, 2003; Barabási; Albert, 1999). From this moment on, social networks became the focus of research. Work has also been carried out on various types of networks to understand their properties and characteristics (Newman, 2003). Based on this it was possible to represent them mathematically, which further boosted the progress of the works that aimed to analyze the characterized networks. Metrics, theories and indices were adopted to measure the behavior of the networks. Work has also been done to different social networks from non-social networks (Newman; Park, 2003).

From the analysis of networks, it is possible to explain several phenomena. Social network analysis allows us to understand the relationship between nodes. Studying these links between nodes for a while raises the question, "How does the evolution of the network occur over time?", understanding the evolution of the network as a whole is a complex task (Al Hasan; Zaki, 2011).

With these concepts in mind, the link prediction problem was proposed (Liben-Nowell; Kleinberg, 2007). Initially, methods were used to calculate the similarity between two network nodes. The more similar the nodes, the more likely they are to be linked together.

Therefore, several other methods have been proposed to solve better the prediction problem of links (Acar *et al*., 2009; Ahmad *et al*., 2020; Kerrache; Alharbi; Benhidour, 2020; Ren *et al*., 2020; Shakibian; Charkari, 2017). Probabilistic, linear algebra-based, and binary classification methods were proposed. Thus, several algorithms can be used for its resolution. This paper will treat links prediction as a classification problem thus, algorithms in the recommendation systems area are used to achieve the proposed objectives.

Applying such concepts to a more specific domain, we can turn our attention to networks belonging to the scientific community. When publishing a paper with another scientist, a connection is formed by the collaboration made. The authors are represented by the nodes and the scientific collaborations or links between them by the edges (Maruyama; Digiampietri, 2019). Such networks are called co-authorship networks and will be our main object of study. According to Rolf (2019), researchers and scientists can improve their decision-making process before getting involved in any project or research group from a more in-depth view of the dynamics that affect scientific collaboration.

In this context, the Lattes Platform, maintained by the National Council for Scientific and Technological Development (*Conselho Nacional de Desenvolvimento Científico e Tecnológico*, CNPq), has been a source of data from various works aimed at analyzing scientific collaboration networks, mainly because it encompasses data from much of the national scientific production (Mena-Chalco; Cesar Junior, 2009; Maruyama; Digiampietri, 2019; Perez-Cervantes, 2015). Lattes Platform currently has 6.6 million researcher curricula and represents one of the world's most relevant and recognized scientific data sources (Lane, 2010). The data in the curricula registered in the Lattes Platform has attributes such as: name, academic background, professional experience, projects, and scientific

publications. The sheer volume of data in curricula can provide valuable and up to now unknown information (Dias *et al.*, 2013).

Understanding the evolution of the network requires understanding how two nodes interact with each other. The relationship between the nodes forms the network, so we seek a way to predict which researchers will produce a joint article in the future. Such behavior is present basically in all social networks through the "suggested friends". Thus, it is possible to use the same techniques for the scientific collaboration networks studied in this work.

The rise of recommendation systems represents a specific approach to machine learning concepts. By employing this technique, it is possible to understand which attributes of the nodes make them closer to each other, and thus have a greater chance of creating a relationship in the future.

Therefore, the prediction of links in co-authoring networks formed by the data of PhDs with curricula registered in the Lattes Platform in the area of Information Science and Biology will be performed and compared. With this, it will be possible to understand the behavior of different knowledge area collaboration networks and monitor its evolution over time. This allows the characteristics that most influence the emergence of new connections to be identified and consequently receive more attention from the scientific community. This study will also determine the researchers who can collaborate in a future instant of time. This information can be helpful for future research projects and strategies to promote research.

The text is organized as follows: Section 2 presents the works related to this and the definition of some important concepts for the execution of the work. Section 3 offers the methods used, explaining all the techniques and decisions taken to complete the job. The results obtained from this methodology will be presented in Section 4. Finally, a conclusion and some future work are presented in Section 5.

Related Work

In a seminal work, the link prediction problem, as we know is defined (Liben-Nowell; Kleinberg, 2007). This study is still considered the starting point for this field. The theme is introduced focusing on social networks and their dynamism. Over time, new edges are added to the networks, which represents the emergence of new interactions in the social structure. The authors define the problem of link prediction as: given a social network at a time t, the goal is to accurately predict the edges that will be added to the network during the interval *t* and a time future *t'*. Link prediction, in this context, allows one to discover individuals who are already working together, but their interaction has not yet been directly observed (Krebs, 2002).

A study aimed to discover which information source could indicate relationships between users (Adamic; Adar, 2003). Throughout this work, several steps are taken to understand one user's connection with another. In this paper, the author refers to the problem as "relationship prediction" and uses a ranking of similar people to predict the missing edges. At the end of the study, a portion of the students was given a list of people most like them, and often recognized as such individuals. The author points out that the great challenge of such analysis is to have only a small data set, which represents a tiny portion of the actual data.

However, to predict a missing link, concepts related to the topological characteristics of the network must be better understood. To this end, a work that focuses on analyzing the main differences between social and non-social networks is conducted (Newman; Park, 2003). It highlighted that the relationship between the degrees of the adjacent nodes of the networks is positively correlated in social networks, but negatively in other types of networks. Secondly, social networks show a high level of clustering. In conclusion, social networks are divided into communities, while non-social networks are not. In this context, we can understand the degrees of a network as the minimum distance, in terms of numbers of areas in the network, between all pairs of nodes in the network, through which a connection exists (Newman, 2001).

Even after several studies in social network analysis, understanding the entire evolution of a network is a complex task, but understanding the association between two specific nodes is much simpler (Al Hasan; Zaki, 2011). Therefore, some questions may be asked: How does the pattern of associations change over time? What are the factors that guide these associations? How is the association between two nodes affected by other nodes? To answer the questions, the author uses the standard problem formulation (Liben-Nowell; Kleinberg, 2007) and surveys existing approaches focusing mainly on social network graphs.

Turning the attention to the networks of scientific collaboration, the object of study of this work, presents one of the first works on this topic. Three specific networks are studied, one in biomedical research, one in physics, and lastly, mathematics. The author presents several characteristics of co-authorship networks and performs several analyses to understand nodes' behavior in this network. The importance of such networks is highlighted, and they have meticulous, well-documented information and even temporal events in scientists' social and professional relationships.

Using the Lattes Platform as a data source, an approach for extracting researchers' curricula and building a scientific collaboration network is described (Dias *et al.*, 2013). The relationship between employees is accomplished through one or more works together. Through the built framework, networks that have standard terms participated in the same congress or even in the same area are presented. In Dias and Moita (2015), the authors present the method in detail. Some tests are performed, and the properties present in them are analyzed.

An approach aiming to find most influential researchers in a collaborative network is presented (Perez-Cervantes *et al*., 2013). For this, a link predictor based on local metrics of the network structure is used. The individual collaborative influence is obtained by considering the influence of a particular researcher on the prediction of network links as a whole. The data from 47,555 Lattes Platform researchers' curricula are used. As a result, the measures of collaborative influence present a significant inverse correlation compared to the most well-known centrality measures. This fact demonstrates the effectiveness of the proposed metrics. Another critical factor is that the described methodology can be calculated independently for each vertex without a global calculation, thus reducing the computational cost (Perez-Cervantes, 2015).

When comparing the present work with the literature review presented above, it is essential to highlight that the seminal works were used as a basis for the research, since they show the problem to be studied. In a second step, articles were used that seek to answer the same questions raised here, but with other databases, to understand the techniques used. Finally, studies using the Lattes Platform went through the same process to better understand the dataset and identify the metrics the authors applied. In this context, the work presented here differs from those previously mentioned in comparing different algorithms, and only two specific areas of knowledge to understand the differences and similarities between their scientific collaboration networks.

Methodological Procedures

To achieve the proposed objectives, some steps are necessary. This section will highlight the methods used to predict future connections in a specific area. First, it is essential to limit the population for the study. For that, knowledge areas were taken into consideration. Initially, due to the proximity of the research group, and previous work, the Information Science area, belonging to the large Applied Social Sciences area, was selected. From the large area of Biological Sciences, Biology was chosen. Researchers with a Ph.D. degree who have their curriculum registered on the Lattes Platform are responsible for 74.51% of published papers in journals, and have a most recent update date for their curriculum. Thus, even though they are a small portion of all those registered, they are the ones who have the best profile to achieve the objectives proposed in work. This data set contains a total of 3,312 Ph.D. researchers' curricula. Initially, the framework used for data extraction will be presented. Secondly, the scientific collaboration networks will be characterized, and lastly, the attributes selected for the prediction will be explained.

To begin the development of the work, it was necessary to perform the extraction of the data to be used. The *LattesDataExplorer* proposed by Dias (2016), a framework for data extraction and processing was used. Initially the data is collected through CNPq and stored in a local repository where data selection is performed. Using the identifier of each curriculum, the date of the last update is compared with the storage in CNPq. If the dates are different, the extractor replaces the curriculum that was stored locally with the most current version (Dias, 2016). Afterward, the data is processed and stored in Extensible Markup Language (XML) format, so that it is possible to generate metrics and calculate some statistics.

With the data extracted and organized, it is necessary to characterize the networks. The co-authorship of an article can be understood as the documentation of a collaboration between two or more authors, and these collaborations form a "network of scientific collaboration" (Newman, 2004). A method for identifying scientific collaborations in large databases using low computational cost was applied to generate the networks used in this work (Dias; Moita, 2015). In this step, articles written together through a search of the entire database are found through the results mentioned above. Since the Lattes Platform does not provide a unique identifier per publication, it is necessary to correlate the researchers' curricula looking for similar information.

After all the resumes are stored in a standard format, the proposed method is applied identifies scientific collaborations. In this method, all the titles of the articles registered in the curriculum of each author are analyzed, and they become the basis of the entire construction of the collaboration network. The steps involved in the identification process are listed in Figure 1.

```
Identification-Collaboration (list-of-publications)

 // Each publication have an id_author
// Co-author bound have an id.
//Each publication is concatenated with the year of publication

  1.  n ← number of articles author
  2.  for i ← 1 to n
  3.    x ← string[i]  // x is article title [i]
  4.    x ← stopword[x]  // removes token without semantic value
  5.    x ← normalization[x]  // remove whitespace and accentuation
  6.    x ← lowercase[x]
  7.    if hash[x] in dictionary  // checks whether x is in the dictionary
  8.      dictionary[x] ← id_author
  9.    else dictionary  ← x, id_author
  10.  return: Adjacency_matriz
```

Figure 1 – Algorithm for identification of collaboration.
Source: Dias and Moita (2015).

As shown in the algorithm for identification of collaboration, each registered title of a study in a particular curriculum undergoes a transformation process that strips the title of accentuation, spaces, and words with no semantic value. The strategic objective of the algorithm is to minimize typos and grammatical errors that may be present in the titles of the articles. Consequently, all the text is standardized in lowercase. The resulting string is concatenated with the year of publication and is subsequently transformed into a key representing the work under review.

Later transformation, the key is inserted in the dictionary that is used for the to characterize of the collaboration network. If the key already exists in the dictionary, the identifier of the originator of the curriculum in

question is linked to the key; otherwise, the key is inserted and becomes an index in the dictionary. Consequently, this dictionary is used to connect the identifiers of each dictionary key to characterize the collaborations of each work, and thus enable the construction of collaboration networks.

After collaboration networks are characterized, it is necessary to identify which attributes will be used for prediction. Therefore, a basic set of features from other works that addressed this theme was selected. The simplest way to perform edge prediction is through the common neighbor's metric (Liben-Nowell; Kleinberg, 2007), which can be understood as the number of common nodes that two specific nodes have. Using this attribute in scientific collaboration networks, it is pointed out that individuals who have never worked together but have a common collaborator are much more likely to collaborate in the future (Newman, 2010). The Common Neighbors (CN) attribute is demonstrated in Eq. 1, where x and y represent vertices of the graph.

$$CN(x,y)=|\Gamma(x)\cap\Gamma(y)| \tag{1}$$

Another metric that can be obtained using the structural characteristics of the network itself is called Jaccard Coefficient (JC), and measures the probability that both x and y have a v neighbor, randomly chosen that x or y own. Unlike the Common Neighbors attribute, the Jaccard Coefficient normalizes the number of common neighbors (Al Hasan; Zaki, 2011), as follows:

$$JC(x,y)= \frac{|\Gamma(x)\cap\Gamma(y)|}{|\Gamma(x)\cup\Gamma(y)|} \tag{2}$$

In order to establish similarity between two pages, Adamic/Adar metric is proposed (Adamic; Adar 2003). In order to use it in link prediction algorithms, it was customized and it is presented in Equation 3 (Liben-Nowell; Kleinberg, 2007). This formulation gives the rarer characteristics a greater weight (Potgieter *et al*., 2009). We can understand it as the number of properties shared by nodes, divided by the log of the frequency of the characteristics.

$$Adamic/Adar(x,y)=\sum_{w\in\Gamma(u)\cap\Gamma(v)} \frac{1}{\log|\Gamma(w)|} \tag{3}$$

Following the same reasoning, the Resource Allocation (RA) metric assigns weight to the two-node relationship favoring relationships between those with few relationships (Digiampietri *et al.*, 2015), and can be found in Equation 4.

$$RA=\sum_{w\in\Gamma(u)\cap\Gamma(v)} \frac{1}{|\Gamma(w)|} \tag{4}$$

Considering only the size of the node neighborhoods, the Preferential Attachment (PA) metric has been proposed and is presented in Equation 5. In short, it establishes that the probability of a new relationship with other vertices is based on the degree of the node in question (Al Hasan; Zaki, 2011).

$$PA=|\Gamma(u)||\Gamma(v)| \tag{5}$$

The fact that friends of friends can create a connection suggests that the distance between nodes in a network can influence the formation of new connections (Al Hasan; Zaki, 2011). In this way, the Shortest Path (SP) metric can also predict links. In graph theory, it would be referred to as the geodesic distance, which can be understood as the shortest path between a pair of nodes (Hoffman; Steinley; Brusco, 2015).

Domain-related attributes can also be used during the prediction process. In this case, it is necessary to evaluate the data set used and the required techniques to convert them to the correct formats for input to the

algorithm. Using the Lattes Platform, various information is present in researchers' curriculum, such as: orientations made, participation in newsstands, congresses in which some publication was held, institutions where the researcher studied, among others. As the Information Science sub-area is already being used in the present work, the fields being used are the city, state, and institution.

While topological attributes are obtained from the execution of some calculations, which use the graph itself, attributes related to the domain are extracted and stored in the same way as the curriculum owner completed it. However, the data must be standardized to facilitate the prediction process when using machine learning techniques. The city, state and institution fields are considered categorical data, and therefore must go through two processes before being used further.

Initially, it is necessary to encode the texts informed by the researcher in numbers. For example, instead of "Belo Horizonte", the value five will be stored for all categorical information; for "São Paulo" the value 13, and so on Label Encoding was applied for this, which is just one of several methods present in Scikit Learn (Buitinck *et al.*, 2013), an open-source machine learning library. Thus, all categorical values were coded in numbers. However, after this process, the algorithms could imply that the value 13, referring to São Paulo, is more important than the value 5, referring to Belo Horizonte. After all, it was not specified that these values represent categories. Therefore, the method called One Hot Encoder, also available in Scikit Learn (Buitinck *et al.*, 2013), must be used. Through it, each category is transformed into a column, and if the value refers to a particular column, the number 1 is inserted, if it is not, a zero is inserted. In this way, categorical data is encoded into a large sparse matrix, composed mostly of zeros.

Finally, the number of collaborations that two nodes had over that time was also considered an attribute. This way it is possible to identify researchers who have been working together longer, and possibly have a more significant influence in the next few moments (Table 1)

Table 1 – Characterized networks.

Area	Period	Researchers	Collaborations	Avg. Degree	Density
Biology	1960 a 2000	2251	367	0.066	0
	2001 a 2010	2251	3446	0.717	0
	2011 a 2019	2251	3983	1.554	0.001
Information Science	1960 a 2000	1061	1064	0.477	0
	2001 a 2010	1061	6084	3.668	0.003
	2011 a 2019	1061	13408	12.535	0.012

Source: Elaborated by the authors (2020).

After defining the attributes that will be used, some steps are necessary. Firstly, it is essential to determine the periods for training and testing, so, for each knowledge area, three different networks were created. For the first network, the publications made between 1960 and 2000, called the initial period, were defined. The second network was created for the period from 2001 to 2010. Finally, 2011 to 2019 was established for the third and last network. Such periods include the date of the first work registered on the platform until the last year before the presentation of this work. Figure 2 shows the six networks generated, where the nodes represent the scientists, and the edges represent papers written together. It is possible to observe that the collaborations between the scientists increased over time from that, and some differences across Biology and Information Science researchers. The objective of the work can be better understood through this figure, where given the first network, the goal is to estimate future networks, that is, to identify the researchers who will work collaboratively in the future.

Table 1 presents the main characteristics of the characterized networks. The amount of edges has increased considerably over the years. It is also important to note that the number of researchers did not change over time.

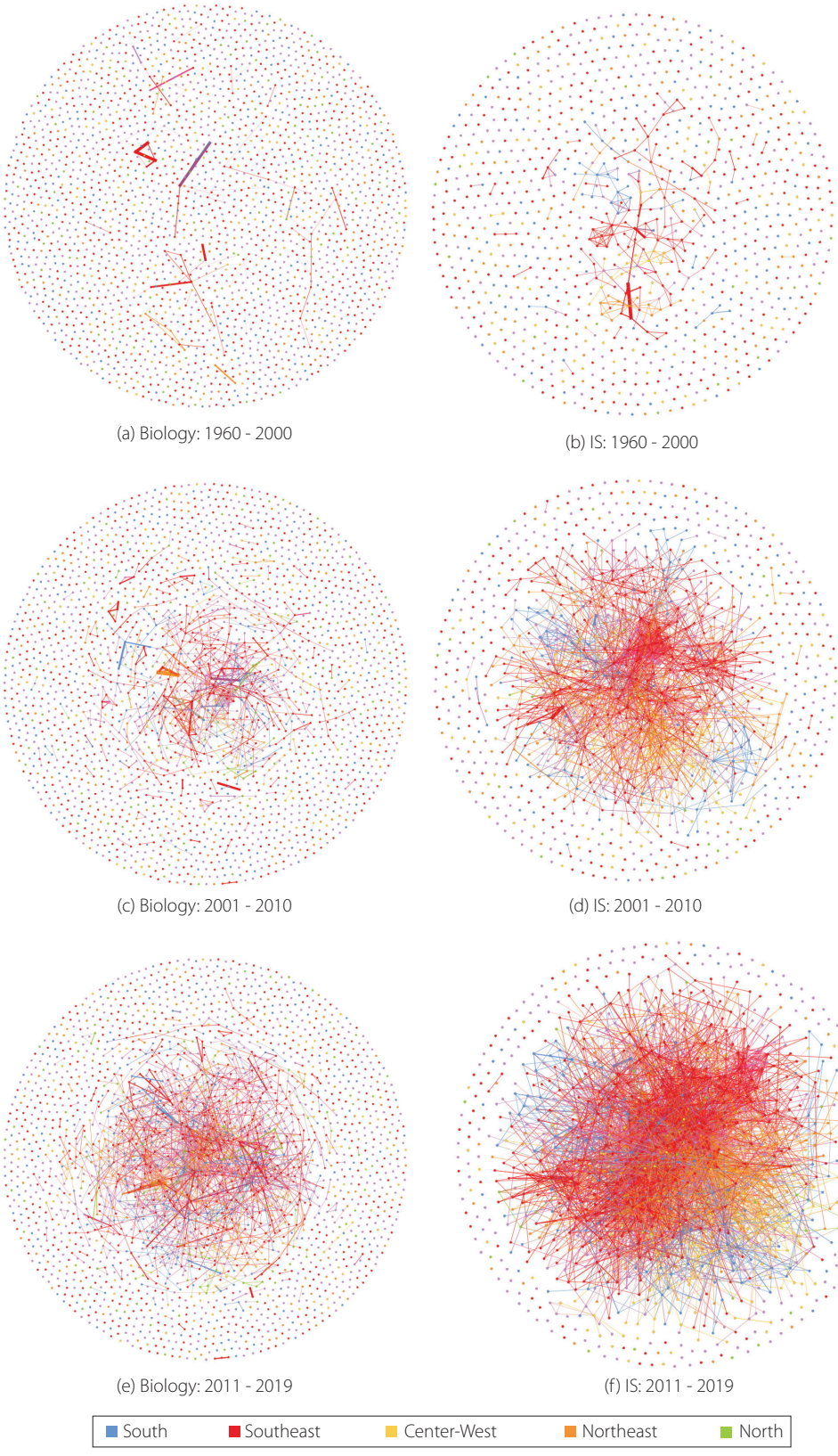

Figure 2 – Network characterization.
Source: Elaborated by the authors (2020).

The same group of nodes was selected for the entire period. The number remains the same because it was not possible to identify when it would be better when researchers should enter the collaboration network. Meaningful collaborations and links could be lost if the date was selected after the Ph.D. conclusion. It would also be necessary to work with targeted collaborations, since a researcher may already be a Ph.D., while his peers are not yet. Thus, all researchers with a Ph.D. degree were considered at the time of data extraction.

A series of information can be obtained when analyzing Biology networks, presented in Figure 2, together with Table 1. It is possible to notice that the network has few scientific collaborations; 367 papers were written together in the first period, increasing to 3,446 in the second period, and ending the analysis period with 3,983 collaborations. This fact can also be observed when analyzing the average degree, and comparing it with the Information Science network, which has less than half the number of researchers, and has more collaborations and, consequently, a higher average degree for all periods.

Taking advantage of the small number of published works in this area, Figure 2 displays some patterns of scientific collaboration. It is possible to notice the presence of more relevant researchers, who connect several other different researchers, even from other regions. Throughout the evolution of the network, it can be seen how new collaborations took place, and the formation of some groups of researchers with more works published together.

Only 1,061 researchers work in Information Science, responsible for over 13,000 works published together, as presented in Table 1. However, an interesting factor of this area, is that even given the smallest number of nodes, it has the highest density, 0.012. It can also be observed, by checking the average degree that this area publishes many papers in collaboration with theirs peers. This value represents that each researcher has published 12.5 papers on average, while this value represents just 1.5 in the Biology area. Also, comparing the number of collaborations, Information Science contains more than three times the number of Biology published papers (Figure 2).

When analyzing the networks considering the country's regions, it is clear that the southeast region is the largest producer of scientific knowledge in both areas. The south results right after the southeast region, followed by the northeast, central-west and north. Such factors can directly influence future scientific collaborations, since it is expected that the most active areas will continue to be the most relevant.

The networks to be used in the rest of the work were presented, and characterized to predict future scientific collaborations. Some topological factors of each network were analyzed; they represent a fundamental factor in the prediction process, since they are essential to calculate some of the attributes used in the algorithms. Researchers' regions were also considered since geographical proximity may represent a fundamental factor in developing joint work.

The researchers' data set, the links between them, and the selected attributes were then used as input to a machine learning algorithm. Each row in the data set is composed of the following items: First Researcher Identification, Second Researcher Identification, Common Neighbors (CN), Jaccard Coefficient (JC), Adamic/Adar (AA), Resource Allocation (RA), Preferential Attachment (PA), Shortest Path (SP), weight, City, State, Institution, and finally the presence or absence of an edge. It is important to note that the indices correspond to the calculations previously presented for the two nodes of the line. The edge is obtained using data from the later period. That is, given this set of attributes, will a new edge be generated? This information will be sent to the prediction algorithm.

At this stage of the work, the problem of class imbalance comes up. The number of possible links in a graph is quadratically related to the number of nodes. However, the number of actual links represents only a small fraction of this number (Al Hasan; Zaki, 2011). This problem interferes with the results due to two reasons: (i) with fewer examples of a given class it is more difficult to infer reliable patterns; (ii) trained models are skewed towards the predominant class (Menon; Elkan, 2011). Several authors propose techniques and methods for solving this challenge (Acar *et al*., 2009; Al Hasan; Zaki, 2011; Perez-Cervantes *et al.*, 2013). A traditional technique for overcoming class imbalance is called under-sampling. It reduces the number of samples of the determinant class randomly,

thus equating the number of components for both cases. This technique was used in work presented here. After under-sampling, the number of edges present and absent is the same. With balanced data, the prediction algorithm was executed.

Results

Throughout the process described in the previous section, the dataset has undergone some changes. The number of positive edges in the network represents only a tiny fraction of the total possible edges. Therefore, it is necessary to go through an under-sampling process. This step is needed to solve the class imbalance problem. If the dataset contained such differences between classes, the prediction algorithms would learn better how to predict non-collaborations than real collaborations, which is the primary goal of this work. It is crucial to clarify that these steps were done for each area separately.

The under-sampling method randomly chooses the same number of absent edges for the dataset to become similar for both classes. Thus, both networks presented the same number of present and missing edges after this procedure, facilitating the prediction process and making the algorithms have the same learning for collaborations and non-collaborations. After this stage, the two sets of data were sliced, where a part was separated from executing the training of the algorithms, and another part for testing. In this way, it is possible to validate a learning experience. Therefore the division was done by selecting 75% of the data for training and 25% for testing (Table 2).

Table 2 – Metrics generated from predictions.

Area	Algorithm	Precision	Recall	F1	AUC
Biology	KNN	0.56	0.55	0.54	0.60
	Logistic Regression	0.73	0.72	0.72	0.79
	Random Forest	0.74	0.72	0.72	0.81
	Naive Bayes	0.66	0.63	0.60	0.61
Information Science	KNN	0.78	0.78	0.78	0.85
	Logistic Regression	0.81	0.81	0.81	0.88
	Random Forest	0.80	0.80	0.80	0.87
	Naive Bayes	0.76	0.69	0.66	0.69

Note: AUC: Area Under the Curve; F1: Weighted average of precision and recall; KNN: K Nearest Neighbor.
Source: Elaborated by the authors (2020).

Several algorithms can be used to solve classification problems. Among them, some were selected to perform the work: Logistic Regression, K Nearest Neighbors, Naive Bayes, and Random Forests. These techniques have a different peculiarity and, consequently, different results. Therefore, their results will be presented in Table 2, using the metrics precision, recall, F1, and Area Under the Curve (AUC). Usually, in link prediction algorithms, the area under the curve is used by most authors to use it as a basis.

Each of the metrics used to validate the results has its characteristics. Accuracy aims to answer the following question: Of all positive predicted values, how many are actually correct? High accuracy is related to fewer false positives. Considering all the positive values, the recall aims to know how many were predicted. The F1 metric takes precision and recall into account, thus making a weighted average of these two metrics. Finally, the AUC is used to present the performance of a classification model throughout the learning process. In practice, AUC calculates the probability that a true link has a higher link prediction score than a non-existing link (Zhang *et al*., 2015).

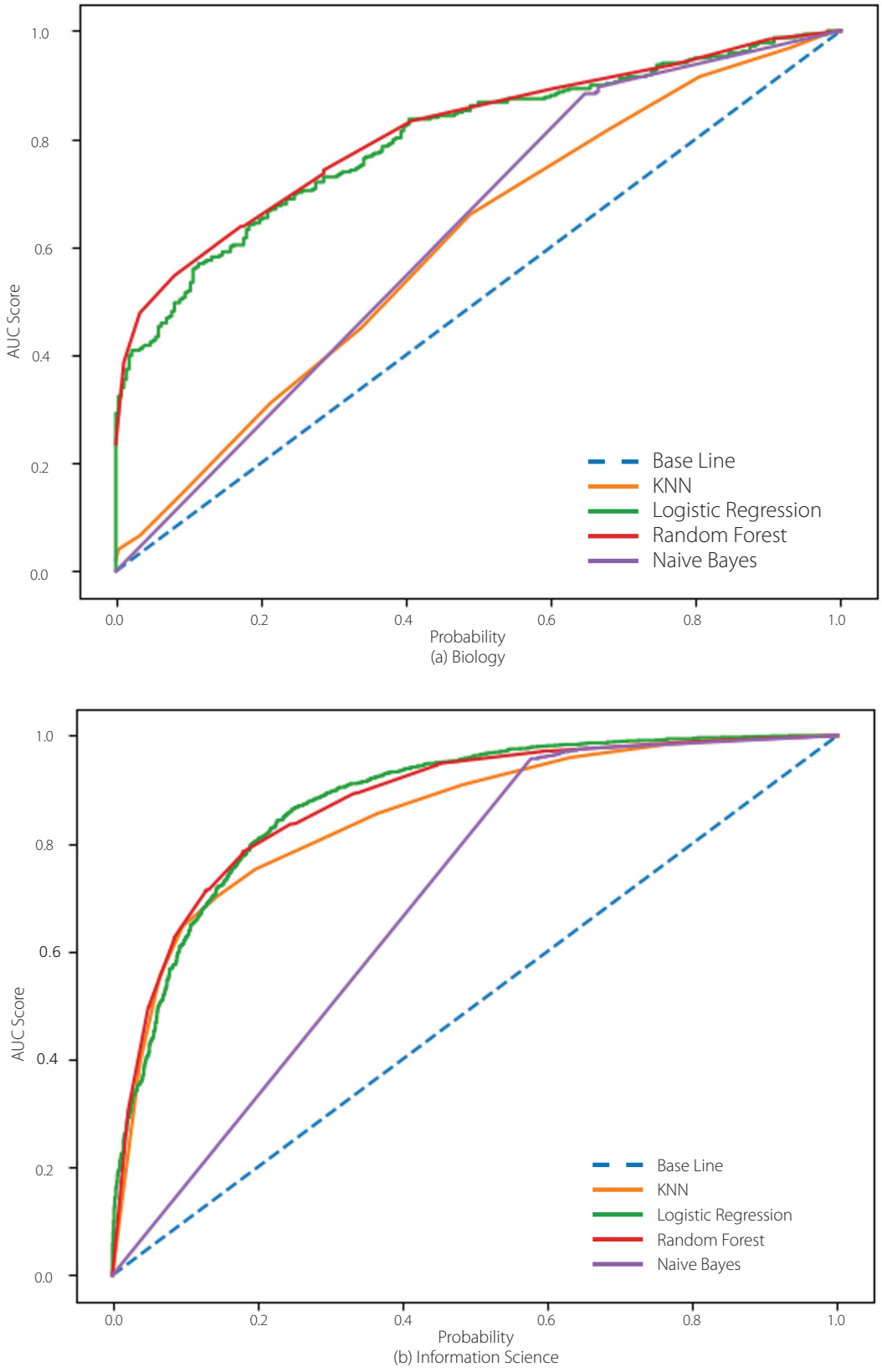

Figure 3 – Area Under the Curve for all algorithms.
Source: Elaborated by the authors (2020).

Analyzing Table 2, it is possible to notice that the chosen algorithms obtained good results. Looking at the area under the curve, we realized that everyone got an impact above a mere chance. This situation is better explained in Figure 3, where the blue dotted line represents a 50% chance of a hit, which means equal odds for the prediction to be correct or incorrect. The orange line represents the values of the predictions made. Thus, the algorithm used the presented data set to make correct predictions about future connections (Figure 3).

Considering the Biology network, the algorithm that presented the best result was Random Forest, with 81.00% of correct answers on the AUC metric, followed by Logistic Regression, Naive Bayes, and K-Nearest Neighbors. The average of all algorithms is 70.25%. However, two algorithms showed results closer to 50%, which can be considered the base case. Considering Information Science network, the best performance, considering all metrics, was the Logistic Regression, followed by Support Vector Machine, Random Forests, K-Nearest Neighbors, and, lastly, Naive Bayes. However, there is a slight difference between the results obtained, making it clear that we cannot yet establish which technique should be used as a standard for the problem in question. In this case, the metrics average is 82.25%, and only the Naive Bayes algorithm showed a result smaller than 70.00%.

The same considerations can be taken from the analysis of Figure 3, which presents the AUC for all algorithms and areas. At this moment, it is possible to realize the importance of the base scenario, established by the blue line. This value represents a mere chance that the prediction was right or wrong since it would be correct 50% of the time and the other 50% of the time, wrong. In this exact figure we can verify the performance of the algorithms and the significant difference between the two areas.

When analyzing the difference between the areas, the algorithms, even following the same methodology, present a better result for Information Science. The results for this area are 12% better than for Biology, demonstrating the great importance of topological attributes during the prediction process. This difference is also seen in Figure 3, where we can consider the learning process over several interactions of the algorithms. That is why it is essential to use this metric when studying link prediction.

By analyzing the learning process taking into account just the topological attributes used (city, state and institution were not taken into account for this evaluation), it is possible to identify the order of influence of each one of them in the final result. From Figure 4, we can observe which of then was most important during the learning

Figure 4 – Feature Importance.
Source: Elaborated by the authors (2020).

https://doi.org/10.1590/2318-0889202234e200033

process. For the Biology network, the order is: Preferential Attachment, Shortest Path, Weight of Collaborations, Resource Allocation, Jaccard Coefficient, Adamic/Adar and Common Neighbors. Considering the Information Science network, we can check the order of attribute importance: Preferential Attachment, Weight of Collaborations, Shortest Path, Jaccard Coefficient, Resource Allocation, Adamic/Adar, and, finally, Common Neighbors. Analyzing Biology network, some topological attributes didn't play a role at all, as is the case of Common Neighbors and Adamic/Adar.

For both networks, these results present a behavior different from most of the theoretical references studied here, where most of the time, the most relevant attribute is the Common Neighbors. However, in the studies performed here, the Preferential Attachment metric is responsible for a good part of the result.

Conclusion

The results presented here show that it is possible to predict of links using information from the network itself. The proposed objective was then achieved by using these data; for example, it is possible to know if two researchers from the area mentioned above will collaborate in a future instant of time. The performance of the evaluation metrics was around 80% representing a good result. It is possible to use the methods presented here to support decision-making when granting scholarships, determining research groups, and promoting researchers. Although the presented methods can be easily applied to similar studies, one of the significant limitations found is the inability to replicate the works found in the literature regarding link prediction since data sets are not public.

Analyzing the difference in the performance of the algorithms for the two areas presented in this work, it is clear the importance of the topological attributes. It is also important to note that machine learning techniques show better results from using a large data set. Therefore, the small number of scientific collaborations in the field of Biology probably influenced the behavior of algorithms at several levels. However, even with an average prediction of 70%, good results and applications can be created using the methodology presented here. Regarding the difference in the most important attributes about other results in the bibliography, it is believed that, because the area of knowledge has already been defined previously, the Common Neighbors attribute may have become less relevant in this analysis.

As future work, we highlight the importance of increasing the data set, or even looking for other ways to solve the class imbalance problem, thus increasing the number of samples present for training the algorithm. From this, the classifiers are expected to perform even better.

Acknowledgment

The authors would like to thank Federal Center for Technological Education of Minas Gerais (CEFET-MG) and Coordination of Superior Level Staff Improvement (Capes) for their assistance in the research.

Contributors

F. AFFONSO was responsible for organizing and structuring the data set, developing machine learning algorithms and generating results. M. O. SANTIAGO was responsible for comparative analysis and validation of the presented results. T. M. RODRIGUES DIAS was research coordinator, supervisor of all stages, from data extraction to generation and analysis of the results presented.

References

Acar, E. *et al*. Link prediction on evolving data using matrix and tensor factorizations. *In*: IEEE International Conference on Data Mining Workshops, 2009, Miami. *Proceedings online* […]. Miami: IEEE Computer Society, 2009. p. 262-269. Doi: https://doi.org/10.1109/ICDMW.2009.5.

Adamic, L. A.; Adar, E. Friends and neighbors on the web. *Social Networks*, v. 25, n. 3, p. 211-230, 2003. Doi: https://doi.org/10.1016/S0378-8733(03)00009-1.

Ahmad, I. *et al.* Missing link prediction using common neighbor and centrality based parameterized algorithm. *Scientific Reports*, v. 10, n. 364, p. 1-10, 2020. Doi: https://doi.org/10.1038/s41598-019-57304-y.

Al Hassan, M.; Zaki, M. J. A survey of link prediction in social networks. *In*: Aggarwal C. (ed.). *Social network data analytics*. Boston: Springer, 2011. p. 243-275. Doi: https://doi.org/10.1007/978-1-4419-8462-3_9.

Barabási, A.; Albert, R. Emergence of scaling in random networks. *Science*, v. 286, n. 5439, p. 509-512, 1999. Doi: https://doi.org/10.1126/science.286.5439.509.

Buitinck, L. *et al.* API design for machine learning software: experiences from the scikit-learn project. ArXiv preprint arXiv:1309.0238, 2013. Available at: https://arxiv.org/pdf/1309.0238.pdf?source=post_elevate_sequence_page. Cited: May 10, 2020.

Dias, T. M. R. *et al*. Modelagem e caracterização de redes científicas: um estudo sobre a Plataforma Lattes. *In*: Brazilian Workshop on Social Network Analysis and Mining (BRASNAM), 2., 2013, Porto Alegre. *Anais eletrônicos* […]. Porto Alegre: Sociedade Brasileira de Computação, 2013. p. 116-121. Available at: https://sol.sbc.org.br/index.php/brasnam/article/view/6851. Cited: May 10, 2020.

Dias, T. M. R. *Um estudo da produção científica brasileira a partir de dados da Plataforma Lattes*. 2016. 181 f. Tese (Doutorado em Modelagem Matemática e Computacional) – Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2016.

Dias, T. M. R.; Moita, G. F. Um método para identificação de colaborações em grandes bases de dados científicos. *Em Questão*, v. 21, n. 2, p. 140-161, 2015. Doi: https://doi.org/10.19132/1808-5245212.140-161.

Digiampietri, L. *et al.* Um sistema de predição de relacionamentos em redes sociais. *In*: Simpósio Brasileiro de Sistemas de Informação (SBSI), 11., 2015, Goiânia. *Anais eletrônicos* […]. Goiânia: Sociedade Brasileira de Computação, 2015. p. 139-146. Doi: https://doi.org/10.5753/sbsi.2015.5810.

Hoffman, M.; Steinley, D.; Brusco, M. J. A note on using the adjusted Rand index for link prediction in networks. *Social Networks*, v. 42, p. 72-79, 2015. Doi: https://doi.org/10.1016/j.socnet.2015.03.002.

Kerrache, S.; Alharbi, R.; Benhidour, H. A Scalable Similarity-popularity Link prediction Method. *Scientific Reports*, v. 10, n. 1, p. 1-14, 2020. Doi: https://doi.org/10.1038/s41598-020-62636-1.

Krebs, V. E. Mapping networks of terrorist cells. *Connections*, v. 24, n. 3, p. 43-52, 2002. Available at: http://ecsocman.hse.ru/data/517/132/1231/mappingterroristnetworks.pdf. Cited: May 10, 2020.

Lane, J. Let's make science metrics more scientific. *Nature*, v. 464, p. 488-489, 2010. Doi: https://doi.org/10.1038/464488a.

Liben-Nowell, D.; Kleinberg, J. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, v. 58, n. 7, p. 1019-1031, 2007. Doi: https://doi.org/10.1002/asi.20591.

Maruyama, W. T.; Digiampietri, L. A. Co-authorship prediction in academic social network. *In*: V Brazilian Workshop on Social Network Analysis and Mining (BRASNAM), 2019, Porto Alegre. *Anais eletrônicos* […]. Porto Alegre: Sociedade Brasileira de Computação, 2019. p. 61-72. Doi: https://doi.org/10.5753/brasnam.2016.6445.

Mena-Chalco, J. P.; Cesar Junior, R. M. Scriptlattes: an open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society*, v. 15, n. 4, p. 31-39, 2009. Doi: https://doi.org/10.1007/BF03194511.

Menon, A. K.; Elkan, C. Link prediction via matrix factorization. *In*: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2011. p. 437-452. Doi: https://doi.org/10.1007/978-3-642-23783-6_28.

Newman, M. E. J. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, v. 101, n. 1, p. 5200-5205, 2004. Doi: https://doi.org/10.1073/pnas.0307545100.

Newman, M. E. J. Mixing patterns in networks. *Physical Review E*, v. 67, n. 2, p. 026126, 2003. Doi: https://doi.org/10.1103/PhysRevE.67.026126.

Newman, M. E. J. *Networks*: an introduction. Oxford: Oxford University Press, 2010. Available at: https://dl.acm.org/doi/book/10.5555/1809753. Accessed on: May 10, 2020.

Newman, M. E. J. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, v. 98, n. 2, p. 404-409, 2001. Doi: https://doi.org/10.1073/pnas.98.2.404.

Newman, M. E. J; Park, J. Why social networks are different from other types of networks. *Physical Review E*, v. 68, n. 3, p. 036122, 2003. Doi: https://doi.org/10.1103/PhysRevE.68.036122.

Perez-Cervantes, E. Análise de redes de colaboração científica: uma abordagem baseada em grafos relacionais com atributos. 2015. Dissertação (Mestrado em Ciência da Computação) –Universidade de São Paulo, São Paulo, 2015. Doi: https://doi.org/10.11606/D.45.2016.tde-18122015-114014.

Perez-Cervantes, E. *et al.* Using Link Prediction to Estimate the Collaborative Influence of Researchers, 2013. *In*: IEEE 9th International Conference on e-Science, 2013, Beijing. *Proceedings online* […]. Beijing: IEEE Computer Society, 2013. p. 293-300. Doi: https://doi.org/10.1109/eScience.2013.32.

Potgieter, A. *et al.* Temporality in link prediction: understanding social complexity. *Emergence*: *Complexity & Organization* (E:CO), v. 11, n. 1, p. 69-83, 2009. Available at: https://aisel.aisnet.org/sprouts_all/195. Cited: May 10, 2020.

Ren, T. *et al.* Identifying vital nodes based on reverse greedy method. *Scientific Reports*, v. 10, n. 1, p. 1-8, 2020. Doi: https://doi.org/10.1038/s41598-020-61722-8.

Rolf, H. Identifying the collaboration styles of research students. *Proceedings of the Association for Information Science and Technology*, v. 56, n. 1, p. 750-751, 2019. Doi: https://doi.org/10.1002/pra2.160.

Shakibian, H.; Charkari, N. M. Mutual information model for link prediction in heterogeneous complex networks. *Scientific Reports*, v. 7, e44981, 2017. Doi: https://doi.org/10.1038/srep44981.

Zhang, P. *et al*. The reconstruction of complex networks with community structure. *Scientific Reports*, v. 5, n. 1, p. 1-11, 2015. Doi: https://doi.org/10.1038/srep17287.