A Strategy for Co-authorship Recommendation: Analysis Using Scientific Data Repositories

Felipe Affonso(✉), Thiago Magela Rodrigues Dias, and Monique de Oliveira Santiago

Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, Brazil
felipe-affonso@hotmail.com, thiagomagela@gmail.com, moniqueosantiago@gmail.com

Abstract. In a co-authorship network papers written together represent the edges, and the authors represent the nodes. By using the concepts of social network analysis, it is possible to better understand the relationship between these nodes. The following question arises: "How does the evolution of the network occur over time?". To answer this question, it is necessary to understand how two nodes interact with one another, that is, what factors are essential for a new connection to be created. The purpose of this paper is to predict connections in co-authorship networks formed by doctors with resumes registered in the Lattes Platform in the area of Information Sciences. To this end, the following steps are performed: initially the data is extracted, later the co-authorship networks are created, then the attributes to be used are defined and calculated, finally the prediction is performed. Currently, the Lattes Platform has 6.1 million resumes from researchers and represents one of the most relevant and recognized scientific repositories worldwide. Through this study, it is possible to understand which attributes of the nodes make them closer to each other, and therefore have a greater chance of creating a connection between them in the future. This work is extremely relevant because it uses a data set that has been little used in previous studies. Through the results it will be possible to establish the evolution of the network of scientific collaborations of researchers at national level, thus helping the development agencies in the selection of future outstanding researchers.

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

1 Introduction

In the late 1990s, several researchers devoted attention to network studies. Work has been done on biology, the internet, routers, among others [4,19,22]. From this

Supported by CAPES.

R. Mugnaini (Ed.): DIONE 2020, LNICST 319, pp. 167–178, 2020.
https://doi.org/10.1007/978-3-030-50072-6_13

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 [20]. 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 differentiate social networks from non-social networks [22].

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 occurs over time?", understanding the evolution of the network as a whole is a complex task [3].

With these concepts in mind, the link prediction problem was proposed [12]. 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 better solve the prediction problem of links [1,14,27]. Probabilistic, linear algebra-based, and binary classification methods were proposed, thus, several algorithms can be used for its resolution. In this paper, we will treat the prediction of links as a classification problem, thus, algorithms in 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. In these networks the authors represent the nodes, and the scientific collaborations represent the edges [16]. Such networks are called co-authorship networks, and will our main object of study.

In this context, the Lattes Platform, maintained by the CNPQ[1], 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. Lattes Platform currently has 6.1 million researcher curricula and represents one of the world's most relevant and recognized scientific data sources [11]. The data in the curricula registered in the Lattes Platform has attributes such as: name, academic background, professional experience, projects, scientific publications, among others. The sheer volume of data in curricula can provide valuable and up to now unknown information [7].

Understanding the evolution of the network requires understanding how two nodes interact with each other. The network is formed by the relationship between the nodes, 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.

Given the arise of recommendation systems, which represent a specific approach to machine learning concepts. By employing this technique it is possible

[1] National Council for Scientific and Technological Development.

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 doctors with curricula registered in the Lattes Platform in the area of Information Science will be performed. With this, it will be possible to understand the behavior of the network and monitor its evolution over time. Through this study, it will also be possible to identify the researchers who can collaborate in a future instant of time. In a second moment, starting from the proposed analysis, it also becomes possible to identify the most influential researchers in the co-authorship network.

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

2 Literature Review

In a seminal work, the link prediction problem, as we know is defined [12]. 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 [10].

With this same goal in mind, a study aiming to discover which source of information could indicate relationships between users [2]. Throughout this work, several steps are taken to understand the connection of one user 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 were given a list of people most similar to them, and often recognized such individuals. The author points out that the great challenge of such analyze is to have only a small data set, which represents a tiny portion of the actual data.

However, in order 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 [22]. It highlighted that the relationship between the degrees of the adjacent nodes of the networks are 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 [19].

Even after several studies in the area of 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 [3]. 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 [12] and conducts a survey of existing approaches focusing mainly on social network graphs.

Turning the attention to the networks of scientific collaboration, object of study of this work, [21] 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 analyzes to understand the behavior of nodes in this network. The importance of such networks is highlighted, and how they have meticulous, well-documented information and even temporal events in the social and professional relationship of scientists.

Using the Lattes Platform as a data source, an approach for extracting researchers' curricula and building a scientific collaboration network is described [7]. The relationship between employees is accomplished through the presence of one or more works together. Through the built framework, networks that have common terms, participated in the same congress or even in the same area are presented. In [6], the authors present the method in detail. Some tests are performed and the properties present in them is analyzed.

An approach aiming to find most influential researchers in a collaborative network is presented [25]. For this, a link predictor based on local metrics of the network structure is used. The collaborative individual influence is obtained by taking into account 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, which were obtained using ScriptLattes [17]. As a result, the measures of collaborative influence present a significant inverse correlation when compared to the most well-known centrality measures. This fact demonstrates the effectiveness of the proposed metrics. Another important factor is that the described methodology can be calculated independently for each vertex, without the need for a global calculation, thus reducing the computational cost [24].

3 Methodology

In order to achieve the proposed objectives, some steps are necessary. This section will highlight the methods used to predict future connections in a specific area. Therefore, the large area of Social and Applied Sciences was chosen, and later the sub-area Information Science. This data set has 1,094 PhD researchers curricula. Initially, the framework used for data extraction will be presented.

Secondly, the scientific collaboration networks created, and lastly, the attributes selected for the prediction will be characterized.

To begin the development of the work, it was necessary to perform the extraction of the data to be used. For this, the *LattesDataExplorer* [8], a framework for data extraction and processing was used. As can be seen in Fig. 1, 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 repository in CNPq. If the dates are different, the extractor replaces the curriculum that was stored locally with the most current version [8]. Afterwards the data is processed and stored in XML (Extensible Markup Language) format, so that it is possible to generate metrics and calculate some statistics.

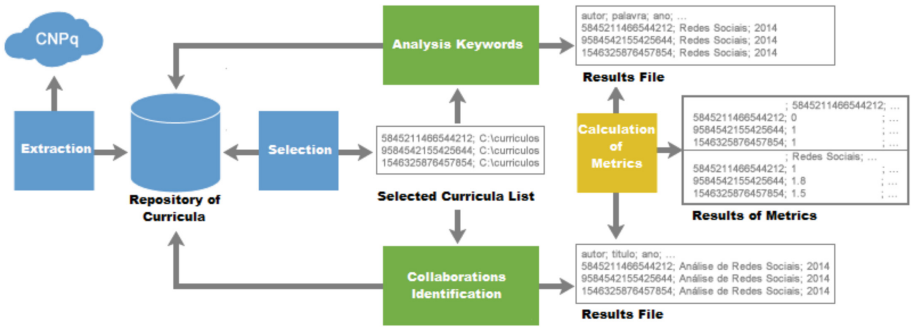

Fig. 1. Framework used for data extraction [8].

With the data extracted and organized, it is necessary to create 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" [21]. A method for identifying scientific collaborations in large databases using low computational cost was used to generate the networks used in this work [5].

After collaboration networks are created, 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 were selected [2,3,9,12,15,24].

The simplest way to perform edge prediction is through the common neighbors metric [12], 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 [23]. 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 [3], 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 [2]. In order to use it in link prediction algorithms, it was customized and it is presented in Eq. 3 [13]. This formulation gives the rarer characteristics a greater weight [26]. 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 [9], and can be found in Eq. 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 Eq. 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 [3]. This metric does not require neighborhood-related information for each node, so it has a lower computational cost [15].

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

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 [3]. In this way, the Shortest Path (SP) metric can also be used in order to predict links. We can understand it as the minimum path between two nodes [9].

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 techniques necessary to convert them to the correct formats for input to the algorithm. By using the Lattes Platform, various information present in the curriculum of researchers, such as: orientations made, participation in newsstands, congresses in which some publication was held, institutions where the researcher studied, among others. In the present work, as the Information Science sub-area is already being used, the data were used: city, state and institution.

The use of categorical data involves a coding process to enable its use in the prediction process. Therefore, each of the aforementioned information was coded

by two processes: LabelEncoding and OneHotEncoder, using libraries already developed for this purpose. From the transformations performed, each categorical data became a sparse matrix.

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

3.1 Proposed Method

After defining the attributes that will be used, some steps are necessary. Firstly it is necessary to define the periods for training and testing, so 3 different networks were created. For network 1, the publications made between 1960 and 2000, which will be called the initial period, were defined. The second network was created for the period from 2001 to 2010. Finally, the period from 2011 to 2018 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 prior to the presentation of this work. Figure 2 presents the 3 networks, it is possible to observe that over time the collaborations between the scientists increased. Through this representation it is also possible to understand the purpose of the work in question. Given the first network, predict what the collaborations will be in the next instant of time.

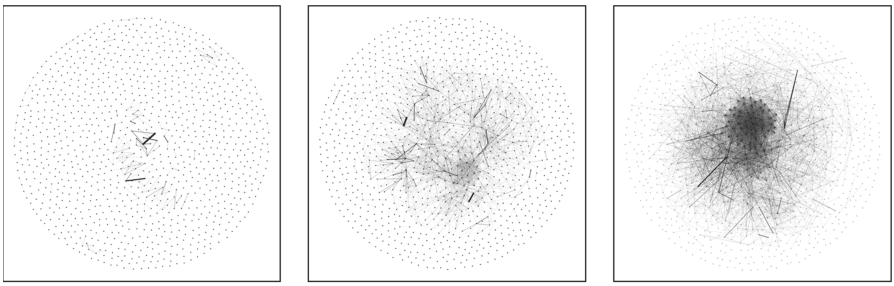

Fig. 2. Evolution of the scientific collaboration network of the information science subarea

Table 1 presents the main characteristics of the created networks. The amount of edges increases considerably over the years. The attributes mentioned above were identified from the use of such networks, most of them use topological attributes to calculate the metric. It is also important to note that the number of researchers did not change over time, the same group of nodes was selected for the entire period.

Table 1. Key networks properties

Network	Period	Number of nodes	Number of edges	Centrality
Network 1	1960–2000	1084	191	0.3524
Network 2	2001–2010	1084	1537	2.8358
Network 3	2011–2018	1084	3831	7.0683

The data set containing the researchers, the links between them and the selected attributes was 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 [3]. 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 [18]. Several authors [1,3,25] propose techniques and methods for solving this challenge. A traditional technique for overcoming class imbalance is called over-sampling. It consists of reducing the number of samples of the determinant class randomly, thus equating the number of components for both cases. This technique was used in the work presented here. Initially the data set had a ratio of 152 missing edges for each edge present. After over-sampling, the number of edges present and absent is the same. With balanced data, the prediction algorithm was executed.

4 Results

Throughout the process described in the previous section, the dataset has undergone some changes. The 1,084 researchers can have a total of 587,528 edges. Of these, only 3,831 represented positive edges in Network 3, so by balancing the samples, a random set of other 3,831 missing edges was chosen. Thus, the data set used for input to the prediction algorithm is made up of 7,662 records. Thus, 5,746 (randomly chosen) lines were selected for the training part, representing 25% of the total set, and another 1,916 were selected for the test part.

Table 2. Metrics generated from predictions

Algorithm	Precision	Recall	F1	AUC
Support Vector Machine	0.78	0.77	0.77	0.86
Logistic Regression	0.78	0.78	0.78	0.87
K-Nearest Neighbors	0.74	0.73	0.73	0.80
Naive Bayes	0.77	0.63	0.58	0.63
Random Forest	0.77	0.77	0.77	0.85

Several algorithms can be used to solve classification problems, among them, some were selected to perform the work, they are: Support Vector Machine, Logistic Regression, Nearest K-Neighbors, Naive Bays and Random Forests. Each of these techniques has 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, so we will use it as basis.

Each of the metrics used to validate the results has its own characteristics. Accuracy aims to answer the following question: Of all positive predicted values, how many are actually correct, high accuracy is related to a fewer false positives. Already considering all the positive values, the recall aims to know how many of these were actually predicted. The F1 metric takes precision and recall into account, thus making a weighted average of these two metrics. Finally, the area under the curve (AUC) is used to present the performance of a classification model throughout the learning process.

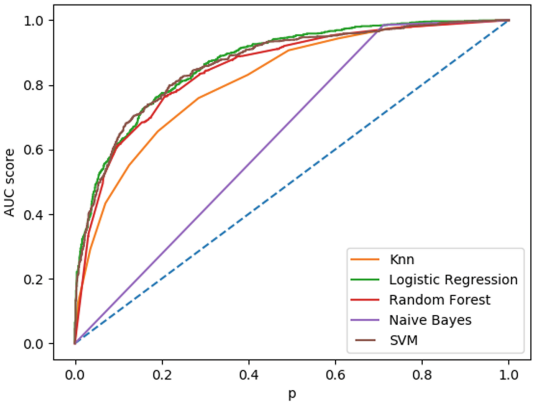

Fig. 3. Area under the curve (AUC)

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 a result above a mere chance. This situation is better explained in Fig. 3, where the blue dotted line represents a 50% chance of hit, which means equal odds for the prediction to be correct or incorrect, and the orange line represents the values of the predictions made. Thus, it is clear that the algorithm was able to use the presented data set to make correct predictions about future connections.

Among the algorithms used, which presented the best performance, taking into account 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, for the problem in question, we cannot yet establish which technique should be used as a standard.

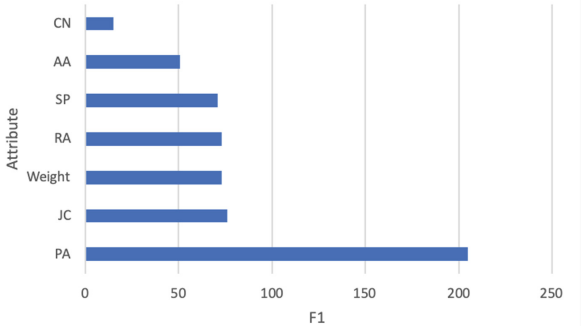

Fig. 4. Feature importance

By analyzing the learning process taking into account the topological attributes used (city, state and institution were not taken in account for this evaluation), it is possible to identify the order of influence of each one of them in the final result. We can observe from Fig. 4 that the order of importance of the attributes for the prediction performed here is: Preferential Attachment, Weight of Collaborations, Shortest Path, Jaccard Coefficient, Resource Allocation, Adamic/Adar, and, finally, Common Neighbors. This fact presents 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 final result.

5 Conclusion

The results presented here show that it is possible to perform the prediction of links using information from the network itself. The proposed objective was then

achieved, since by using these data it is possible to know, for example, 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, but higher values can be achieved by using more attributes, mainly domain-related ones. 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 major limitations found is the inability to replicate the works found in the literature regarding link prediction, since data sets are not public.

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

References

1. Acar, E., Dunlavy, D.M., Kolda, T.G.: Link prediction on evolving data using matrix and tensor factorizations. In: ICDMW 2009. IEEE International Conference on Data Mining Workshops, 2009, pp. 262–269. IEEE (2009)
2. Adamic, L.A., Adar, E.: Friends and neighbors on the web. Soc. Netw. **25**(3), 211–230 (2003)
3. Al Hasan, M., Zaki, M.J.: A survey of link prediction in social networks. In: Social Network Data Analytics, pp. 243–275. Springer (2011). https://doi.org/10.1007/978-1-4419-8462-3_9
4. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
5. Dias, T.M.R., Moita, G.F.: A method for the identification of collaboration in large scientific databases. Em Questão **21**(2), 140–161 (2015)
6. Dias, T.M.R., Moita, G.F., Dias, P.M., Moreira, T.H.J.: Identificação e caracterização de redes científicas de dados currículares. iSys-Revista Brasileira de Sistemas de Informação **7**(3), 5–18 (2014)
7. Dias, T.M., Moita, G.F., Dias, P.M., Moreira, T., Santos, L.: Modelagem e caracterização de redes científicas: um estudo sobre a plataforma lattes. In: BRASNAM-II Brazilian Workshop on Social Network Analysis and Mining, pp. 10–20 (2013)
8. Dias, T.: Um estudo da produção científica brasileira a partir de dados da plataforma lattes. 181p. Programa de Pós-Graduação em Modelagem Matemática e Computacional, Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte (Doutorado) (2016)
9. Digiampietri, L., Maruyama, W.T., Santiago, C., da Silva Lima, J.J.: Um sistema de predição de relacionamentos em redes sociais. In: Brazilian Symposium on Information Systems, vol. 11 (2015)
10. Krebs, V.E.: Mapping networks of terrorist cells. Connections **24**(3), 43–52 (2002)
11. Lane, J.: Let's make science metrics more scientific. Nature **464**(7288), 488 (2010)
12. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. In: Conference on Information and Knowledge Management (CIKM 2003), pp. 556–559 (2003)

13. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. Technol. **58**(7), 1019–1031 (2007)
14. Liu, Z., Zhang, Q.M., Lü, L., Zhou, T.: Link prediction in complex networks: a local Naïve Bayes model. EPL (Europhys. Lett.) **96**(4), 48007 (2011)
15. Lü, L., Zhou, T.: Link prediction in complex networks: a survey. Phys.: Stat. Mech. Appl. **390**(6), 1150–1170 (2011)
16. Maruyama, W.T., Digiampietri, L.A.: Co-authorship prediction in academic social network. In: Anais do V Workshop Brasileiro de Análise de Redes Sociais e Mineração, pp. 79–90. SBC (2019)
17. Mena-Chalco, J.P., Junior, R.M.C.: Scriptlattes: an open-source knowledge extraction system from the lattes platform. J. Braz. Comput. Soc. **15**(4), 31–39 (2009)
18. Menon, A.K., Elkan, C.: Link prediction via matrix factorization. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011. LNCS (LNAI), vol. 6912, pp. 437–452. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23783-6_28
19. Newman, M.E.: The structure of scientific collaboration networks. Proc. Nat. Acad. Sci. **98**(2), 404–409 (2001)
20. Newman, M.E.: Mixing patterns in networks. Phys. Rev. E **67**(2), 026126 (2003)
21. Newman, M.E.: Coauthorship networks and patterns of scientific collaboration. Proc. Nat. Acad. Sci. **101**(suppl 1), 5200–5205 (2004)
22. Newman, M.E., Park, J.: Why social networks are different from other types of networks. Phys. Rev. E **68**(3), 036122 (2003)
23. Newman, M.: Networks: An introduction. Oxford University Press, Oxford (2010)
24. Perez Cervantes, E.: Análise de redes de colaboração científica: uma abordagem baseada em grafos relacionais com atributos. Ph.D. thesis, Universidade de São Paulo (2015)
25. Perez-Cervantes, E., Mena-Chalco, J.P., De Oliveira, M.C.F., Cesar, R.M.: Using link prediction to estimate the collaborative influence of researchers. In: 2013 IEEE 9th International Conference on eScience (eScience), pp. 293–300. IEEE (2013)
26. Potgieter, A., April, K.A., Cooke, R.J., Osunmakinde, I.O.: Temporality in link prediction: understanding social complexity. Emerg.: Complex. Organ. (E: CO) **11**(1), 69–83 (2009)
27. Zhou, T., Lü, L., Zhang, Y.C.: Predicting missing links via local information. Eur. Phys. J. B **71**(4), 623–630 (2009)